RICE UNIVERSITY

The effect of secondary tasks and stimulus type on ratings of telephone hold workload

by

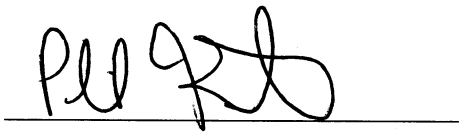
Andy Su

A THESIS SUBMITTED
IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE

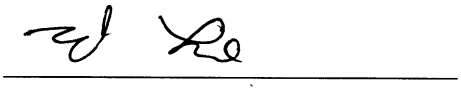
Master of Arts

Approved, Thesis Committee:

Mike Byrne, Professor, Psychology



Phil Kortum, Professor, Psychology



David Lane, Professor, Psychology

Houston Texas

February 2011

ABSTRACT

The effect of secondary tasks and stimulus type on ratings of telephone hold workload

By

Andy Su

Auditory progress bars (APBs) are aural stimuli designed to convey time progression. To investigate the relationship of APBs and workload ratings during a telephone holding context, two APBs were tested alongside ethnographically-validated caller secondary behaviors in a multitasking procedure. Predictions based on Multiple Resources Theory were found to be absent or in the opposite direction, in that an electronic musical APB was rated higher than a voice-based APB in workload as measured by NASA-TLX and task performance. Differences between APBs were manifest through both overall workload ratings and NASA-TLX subcomponent scores. Results indicate that workload measurement can be noisy, particularly when task demands are low to moderate, and that the small effect of APB type may be less important than other considerations for APB design.

ACKNOWLEDGEMENTS

Thanks to:

Dr. Phil Kortum, advisor

Dr. Mike Byrne and Dr. David Lane, committee members

Mr. Ronald Carmona for helping collect some of the data

Mr. Sebastian Thomas for his support

Table of Contents

List of Figures and Tables

INTRODUCTION

Despite the availability of newer technologies, the telephone is still a primary means of obtaining customer service for a large segment of the population. The International Customer Management Institute has reported that while the telephone is more costly than web-based service solutions, it continues to be the more frequently accessed (ICMI, 2008). Similarly, a survey of wireless subscribers in 2006 found that 73% of those polled sought customer service through the phone, compared with only 3% that used the Internet (JD Powers & Assoc., 2006). In order to cut costs, hold queues are used by telephony customer service providers to handle large volumes of calls with fewer operators. Thus it is almost inevitable that most people will experience waiting on hold when they seek customer service. Waiting, whether in a line or on hold, is typically a dissatisfying experience for customers (Munichor & Rafaeli, 2007). Auditory progress bars (APBs) aim to increase customer satisfaction by providing them with a means of accurately estimating hold time. Auditory progress bars are the sonic equivalent of visual progress bars, which are now ubiquitous in a variety of computer applications, providing users with a visible indication of time to task completion for such activities as file transfers, program installations, or virtually any task that takes more than a few seconds. Auditory progress bars aim to perform the same type of service for sound-driven interfaces, such as interactive voice response systems. By giving them a sense of the time elapsed and time remaining in the queue, APBs will ideally enable customers to multi-task efficiently while on hold and make a fair assessment of the quality of the service provider. To facilitate caller multitasking while on hold, APBs need to make minimal demands on the attention of the caller while still conveying the temporal information. Alternately, it may be said that a major goal of APB design is to minimize the mental workload required for callers to use them effectively. The proposed studies make the following investigations:

1. To extend and validate previous work in establishing and describing caller multitasking behavior in a naturalistic setting, and provide representative secondary tasks for laboratory study.
2. To measure the baseline mental workload of the representative tasks found in Part 1 as benchmarks against which APB workloads can be compared.
3. To investigate the mental workload of several types of APBs as well as other on-hold stimuli, to learn how best to minimize the attentional requirements.

REVIEW OF RELEVANT LITERATURE

Time perception

Auditory progress bars operate by leveraging a number of cognitive and perceptive mechanisms which serve to influence our perception of time. The relationship between judgments of duration length and the number and complexity of stimuli present is moderated by whether the estimator is required to process external information (Hicks, Miller, & Kinsbourne, 1976). When the time estimation task is the sole focus of the estimator, filled intervals are usually judged to be longer than equivalent lengths of silence (Fraisse, 1984). This is also known as the filled-duration illusion, and has been replicated for both the visual and auditory domains (Thomas & Brown, 1974, Thomas & Brown, 1975). Meanwhile, when given a secondary task, the direction of the relationship is inverted and participants judged more complex stimuli to be shorter (Hicks, Miller, & Kinsbourne, 1976). In the context of mental workload for APBs, this suggests that the mere presence of any stimuli affects the amount of processing done, as evidenced by the change in time perception.

There also exists a division between experienced and remembered time perception. When subjects are aware that a time judgment must be made, and are actively attending to duration information they are said to be judging time prospectively. When subjects are not made aware of the duration judgment task and are asked later to recall how long a duration was, it is referred to as a retrospective judgment. There is evidence that different mechanisms underlie these two modes of time experience, with an attentional account of prospective judgments and a memory-based model of retrospective judgments (Block & Zakay, 1997). Under an executive function cost incurred by the demands of a secondary task, prospective judgments tend to decrease while retrospective ones tend to increase. As fewer attentional resources can be allocated to the time perception task due to the secondary task, prospective judgments are inversely related to the executive function cost (Block & Zakay, 2004). For the reconstructive process of retrospective estimation, more numerous and complex stimuli during the interval will lead to longer estimates, since more events usually take more time to occur (Kellaris & Mantel, 2003). In the customer satisfaction context, both prospective and retrospective judgments of hold time are significant, as each contributes to the customer's overall evaluation of the quality of service, both during and after the service has been provided.

A host of other factors can interact in a complex way to affect time perception and duration judgment, making manipulation of time perception a nontrivial pursuit. Increased attention tends to lengthen duration judgments (Tekman 1997, Brigner 1988, Meyer, Shinar, & Leiser 1990). Arousal levels and emotional valence can also impact temporal perception, where heightened physical arousal leads to shortened time perception (Jamin et al. 2004). These and other elements are manipulated in aural signals that become APBs, and it is reasonable to think that besides an effect on time estimation, they affect mental workload as well. Therefore, care needs to be taken when considering the ways in which mental workload of APBs can be minimized.

Waiting and satisfaction

There exists a large body of work regarding perception of wait times beyond the scope of psychological laboratory research, in the marketing and customer service context. When considering factors to influence telephone on-hold stimuli and the perception of call wait queues, these data can be informative. As with laboratory time perception, customers' perception of wait times is rarely accurate and has a large impact on their perception of the quality of service offered.

It is perhaps unsurprising that people have a general tendency to retrospectively overestimate the amount of time that they have been waiting. Indeed, Hornik found that shoppers in line at supermarkets and banks overestimated the length of their wait in line by over 30% (1984). Passengers waiting for the bus were found to overestimate their wait time by an average of 0.84 minutes (Mishalani & McCord, 2006). Jones and Peppiatt (1996) found similar data for shoppers in fast food restaurants, with overestimation ranging from 36% to 40%. It was noted that the shorter the actual wait time, the more severe the overestimation became.

For callers on hold, music is a frequent stimulus, yet the role music plays in wait time perception is often quite unexpected. The presence of music alone does not necessarily predict whether a customer will over or under-estimate the wait. Rather, the emotional valance of the music has an effect, with positively valanced music appearing to result in longer estimations than negatively valanced music (Kellaris & Kent, 1992; Hui, Dube, & Chebat, 1995). However, these longer estimates do not necessarily lead to unfavorable ratings of customer service. Indeed, the mere presence of music can make people more willing to wait (North & Hargreaves, 1999). Thus, satisfaction about a waiting experience is influenced by more than just the amount of perceived and actual waiting time, though these do play a large role. The consumer's mood, expectations, and cost incurred all interact to produce the ultimate perception of satisfactory or inadequate service (North, Hargreaves, & McKendrick, 1999; Antonides, Verhoef, & Aalst, 2002; Cameron, Baker, Peterson, & Braunsberger, 2003). It is often

difficult to predict which factors are best leveraged to influence customer perceptions of wait time and service quality. Indeed some factors, such as customer mood, are not easily manipulated in an applied setting. Since the ultimate goal of auditory progress bar research is to produce working and effective implementations, how APBs influence these interconnected factors merits closer examination.

Previous work in auditory progress bars

The current experiments continue a series of studies on APBs and seek to replicate some of the previous findings. Polkolski and Lewis (2002) changed the rate of ticking of audible tones and found that as ticking rate increased, participants' estimates of the waiting period also increased. This was an important precedent for subsequent research in the manipulation of wait time perception. Crease and Brewster (1998) coined the term APB in their investigations of pairing auditory signals with visual progress indicators. In another precursor to current APB research, Knott, Kortum, Bushey, and Bias (2004) examined the effect of music choice and announcement duration on subjective wait times. While not strictly an experiment on APBs, the results from that experiment show that is it possible to manage customer perceptions of hold time. By changing the length of the brief message at the beginning of service calls, the experimenters were able to manipulate the time attribution of customers. When the message is long, the time taken listening to the message is regarded by callers as active time as opposed to hold time, and the subjects in this condition estimated hold time more accurately than callers in the short message condition, who overestimated significantly.

In order to influence perceived hold time in a more controlled fashion, Kortum et al. (2005) introduced auditory progress bar consisting of computer generated tones, which were used in isolation without being paired with any visual aids. These tonal APBs varied along two dimensions - the auditory property being changed (pitch or duration), and the direction of change (increasing or decreasing). While no significant main effects were found, there was a strong interaction effect

between direction and dimension. Therefore the data suggested that certain combinations of factors could be more effective in aiding estimation than others. However, customer satisfaction was generally inadequate for all the APBs in this study, perhaps due to their simple and unglamorous construction.

Seeking to remedy the low satisfaction scores, Kortum et al. (2006) turned to whole pieces of music as APBs. The callers were informed that when the piece of music they heard ended, they would be able to speak to a live agent. This differentiates whole-song APBs from normal hold music, which conveys no temporal information. The whole-song APBs proved to be equally effective in aiding estimation as the best tonal APBs from the previous study, in addition to producing substantially higher customer satisfaction. However, whole songs are difficult to obtain for very long and short durations, and rely on the listener knowing the song in order to provide accurate queue information. When the listeners are unfamiliar with the song, it becomes much less effective as a progress bar. These and other issues prevent whole songs from being practical working APBs. More recently, Kortum, Ling, Su, Peres, and Stallman (2008) investigated the effect of APB type and secondary task on subjective workload ratings. The types of APBs investigated included voice prompts, sinusoidal tones, and electronically generated musical sequences. It was found that the APB type and secondary task do influence perceived mental workload, but their relationship is complex. It appears that in addition to the type of stimuli and secondary task, the simple act of being on the phone contributes to perceived workload. Therefore, the following studies were designed to isolate this effect.

Mental workload

Workload is a somewhat unclear concept in the field of human factors. While few would argue that past some threshold, increasing task demands on human operators begins to decrease performance, the specific mechanisms of the processes involved are points of contention. While task demands can place workload on the operator in both the physical (force exertion) and mental (cognitive effort)

capacities, it is primarily mental workload which is ill-defined and without consensus. The theoretical frameworks which characterize and define mental workload are rooted in models of attention, while operational definitions of mental workload are as varied as the experimental paradigms used to investigate it. Therefore, it becomes necessary to choose a framework and operational definition in any investigation involving mental workload.

How mental workload is viewed, and consequently measured, can be a function of whether it is viewed as a unitary or fragmented resource. The single-channel view contends that cognitive processing occurs serially, and therefore one task must be completed to free up resources for the next (Welford, 1952; Welford, 1967). Mental workload then manifests when there are more tasks than available resources, causing a bottleneck. The single-channel view has been challenged, however, on the basis that there can sometimes be savings for multiple concurrent tasks, often with separate visual and auditory components (Allport, 1980; Allport, Antonis, & Reynolds, 1972; Wickens, Sandry, & Vidulich, 1983). Consequently, models which can support the processing of concurrent tasks at least partly in parallel were developed (Navon & Gopher, 1979; Navon, 1984, Wickens, 1984a; Wickens, 1984b). Wickens's Multiple Resource Model is a well-known framework which suggests that disparate resources are drawn upon in order to perform tasks, and tasks which do not require the same resources can operate in parallel with little cost. This in essence cannibalizes the single-bottleneck model, as it shares the same assumption that it is competition for resources which drives performance (and by extension workload), instead of cross-resource interference or some other process (Navon & Miller, 2002). For the current studies, a multiple resources framework is more fitting, due to the dual-tasking by task type design.

Mental workload can be operationally defined in a number of ways. Workload measures generally fall into three categories: physiological, performance, and subjective. Physiological measures, such as heart rate, assume that increased resource expenditure can be detected through physical

activations. Performance measures score the level of accomplishment of primary or secondary tasks and assume that increased resource expenditure leads to decreased performance. Subjective measurements ask subjects to estimate their own experience of workload on one or more dimensions, and assume that people are generally capable of providing such information.

Each of these types of measurements has its benefits and drawbacks. Physiological measures are objective, continuous, time-sensitive, and can be measured in the absence of behavior (Damos, 1991). However, they can be intrusive depending on the measure and the task, and have significant barriers to entry in the form of dedicated equipment and operator training. Furthermore, physiological measures can decrease the realism of any experiment seeking to replicate real user behavior in actual use contexts. Due to these drawbacks, physiological measures were omitted for the present studies.

Performance measures, especially of the dual-task variety, can have higher diagnosticity to isolate the sources of mental workload (Waard, 2005). Performance measures can also be highly generalizeable to real-world task performance if experimental tasks are analogous to those in the field (Scribner, Wiley, Harper & Kelley, 2007; Wickens, Goh, Helleberg, Horrey, & Talleur, 2003). Meanwhile, the relationship between performance and workload is complex. It is not always the case that as task demands increase, performance decreases. Increased effort and use of sophisticated strategy can lead to the preservation of performance even as demands increase. Similarly, very low workload can lead to boredom and underperformance (Nachreiner, 1995; Sawin & Scerbo, 1995). In these cases, the changes in associated workload would be invisible to performance measures alone. Straightforward primary and secondary task performance measures were included in the present experiments.

Hart and Staveland (1988) elaborated on the role of motivation, operator expectation, and desired level of performance on workload, and cite those factors as being the reason for the inclusion of several sub-scales in their subjective workload measure. Subjective workload measures are easy to implement, can be single or multi-dimensional, and have been claimed to be well-validated (though

some would dispute this), leading to suggestions that they be the criteria against which objective performance measures should be calibrated (Jex, 1988). Jex's operationalization of mental workload (1988, p11) is a useful one for the current experiments:

> "Mental workload is the operator's evaluation of the attentional load margin (between their motivated capacity and the current task demands) while achieving adequate task performance in a mission-relevant context."

There are numerous scales for subjective measurement of mental workload. Some of the more widely known are the Cooper-Harper Scale (Cooper & Harper, 1969), the Bedford Scale (Roscoe, 1987; Roscoe & Ellis, 1990), the SWAT (Subjective Workload Assessment Technique) (Reid & Nygren, 1988), and the NASA-TLX (Task Load Index) (Hart & Staveland, 1988). The SWAT and the NASA-TLX are well-established, and in particular the NASA-TLX has been used and validated in a wide range of situations. It has been shown to correlate highly with time estimation performance (Lind & Sundvall, 2007), have higher sensitivity than the SWAT (Rubio, Diaz, Martin, & Puente, 2004), and be sensitive to task difficulty in flight simulation environments (Selcon, Taylor, & Koritsas, 1991). In a review of 550 experiments using the NASA-TLX, Hart (2006) remarks that the NASA-TLX has become a standard against which newer measures are often benchmarked. In addition to being well-established and validated, the NASA-TLX also has a significant advantage in providing six subscales, which can be decomposed to provide a more detailed description of how participants experience task load (Hart, 2006). These characteristics made the NASA-TLX a good tool for the subjective portion of the workload measurement in the current experiments. Additionally, past experiments in this series have also used the NASA-TLX, so continued use of that instrument is necessary to enable comparison.

By combining these task performance and subjective measures of workload, the present studies compare the demands that APBs place on users in a multitasking setting and identify the best APB candidate to satisfy the stated goals of facilitating caller multitasking behavior.

Part One – Ethnographic study of telephone user behavior

OVERVIEW

Previous research has suggested that users multi-task while waiting on telephone hold (Kortum & Peres, 2007). However, due to the self-report nature of the data, the nature, duration, and frequency of the tasks are not well established. Furthermore, it is unknown whether people engage in more than one secondary task per call, or if any task switching takes place. To accurately gage the impact of auditory stimuli on multitasking, a representative selection of secondary tasks must be found. These tasks should be faithful to natural user behavior, and diverse so as to cover a spectrum of modalities and cognitive demands.

An ethnographic study was conducted to observe users' hold behavior in their own homes. The use of ethnography in human factors research is well established. Millen (2000) introduced a number of strategies for rapid ethnographic techniques to quickly gather user behaviors at low cost, extending earlier techniques outlined by Anderson (1992). Coleman, Hand, Macaulay, and Newell (2005) applied ethnographic methods to research auditory interface design processes, using interviews and observation to gather developer behaviors in their natural work environment. Ethnographic methods allow the observation of users outside of the laboratory in order to capture naturalistic behaviors, as many secondary tasks are simply not available or feasible in the lab. For example, while at home an idle computer may be an invitation to surf the web, participants in the lab may be reluctant to tinker with lab equipment for fear of doing something inappropriate. In fact, this exact scenario occurred during pilot testing, even when the experimenter hinted at the availability of the computer for use by leaving a web browser pointed at Google on the screen. The ethnographic design of this study is intended to capture the broadest range of behaviors possible. The captured behaviors were cataloged, and a representative sampling of tasks was selected for use in Part 3.

METHOD

The ethnography was conducted via remote recording under the guise of a study on teleconferencing. A digital video recorder was placed in the participant's home by the experimenter. The subject was then asked to make three calls to a computerized interactive voice response system as a part of the cover story, and placed on hold each time. The footage of the users' behavior during the telephone call was then analyzed to extract the relevant behavioral information.

Participants

A total of 38 participants were recruited from the Rice University undergraduate population, consisting of 26 males and 12 females and with an age range of 18-22. All participants were screened for normal or corrected-to-normal hearing and vision, and received credit toward a course requirement as incentive.

Materials

The interactive voice response system that the participants call into was constructed with Pronexus Software's VBVoice development platform for Microsoft Visual Basic. Its main role was to provide a plausible cover story with which to keep the subjects on hold so their multitasking behaviors can be captured. An American male voice synthesizer from AT&T's Natural Voices Text to Speech Demo was chosen to be the IVR personality. This voice was chosen over a live voice because it had the characteristics of being intelligible yet clearly computerized. It was hoped that this would maximize the callers' impatience with the system and induce multitasking. The IVR ran on a Dell Opteron PC with an Intel Dialogic voice card connected to a standard phone line. Due to hardware limitations, only one call was processed at once.

Upon answering the phone and greeting the caller, the system reminded the caller to make sure

that the camera is recording. With confirmation, the caller was asked to choose the number corresponding to the call made (1, 2, or 3), at which time the IVR told the caller it must take time to assign the caller to the proper experimental condition, and placed the caller on hold for 210 seconds. The caller was told at the beginning of the hold period that they would have to wait anywhere from one to five minutes. This provided the caller with the knowledge that the process may take a while, and thus encouraged switching to a secondary task. The time estimate was left intentionally vague to prevent callers from abandoning the phone altogether and coming back in five minutes. During the hold period light music or "muzak" was played, selected to mimic many commercial IVRs. At the end of the hold period, the system assigned a different cover task for each call. All of the cover tasks involved making gestures toward the camera, in keeping with the deceptive cover story. For example, participants were asked to convey anger to the camera, or to gesture for the other party to reply via e-mail.

Two cameras were used for the study, a StarTech ST-DVR063 self-contained digital video recorder, and a Sony Cyber-shot digital camera with video capability. Both cameras were attached to small, freestanding tripods. Participants had full knowledge of the camera's presence, as well as full control over the recording function. They were asked to begin recording before each call and end the recording after the completion of each call. This both protected the participants' privacy, and reduced the amount of video that must be analyzed.

Procedure

Due to the nature of the protocol, participants who signed up were asked to contact the experimenter to set up an appointment time. During the meeting, the experimenter obtained informed consent from the participant, gave an overview of the study, and provided training on usage of the camera. The experimenter then accompanied the participant back to his/her home in order to install the camera. The participants were given detailed printed instructions on each step of the study, as well as

reference materials for camera operation.

Participants had 24 hours to place a total of three calls into the IVR system. They were instructed to make no more than one call per hour, in order to make sure that they would be performing routine daily tasks before and after each call. In case of a disconnection, they were asked to wait five minutes before attempting to call again in order to allow the system to reset.

Participants were instructed to turn on the camera and begin recording before each call began. They then called the experimental number and were moved through the IVR system. Afterward, the participants were asked to turn off the camera.

Participants were debriefed at an appointed time one day after completion of the experiment. The experimenter reviewed the captured video footage with the participants, in order to clarify ambiguous behavior and establish context of use. Participants also completed a self-report questionnaire similar to the one used by Kortum and Peres (2007). This provided further subjective data on the way users multitask during hold time, and could compared with the objective data obtained via video to reveal any inconsistencies between what participants report and what they actually did.

RESULTS

The video data were reduced through a time-task analysis, with a temporal granularity of 10 seconds. A scoring rubric was developed that operationally defined tasks of interest, which were based on those from Kortum and Peres (2007). All behaviors were categorized into one of the activities shown in Figure 1. Each occurrence of a secondary task was tabulated for frequency and time – in other words, how often it happened, how long it took, and at what point during the call. Listed by frequency, the top five observed secondary behaviors during the hold period were Nothing, Web Browsing, E-mail, Homework, and Item Manipulation. The least frequently observed behaviors were

Eating/Drinking, Non-calling Phone Use, Active Listening, Talking with Others, and Computer/Video Gaming. Table 1 reports the total task times for all recorded activities.

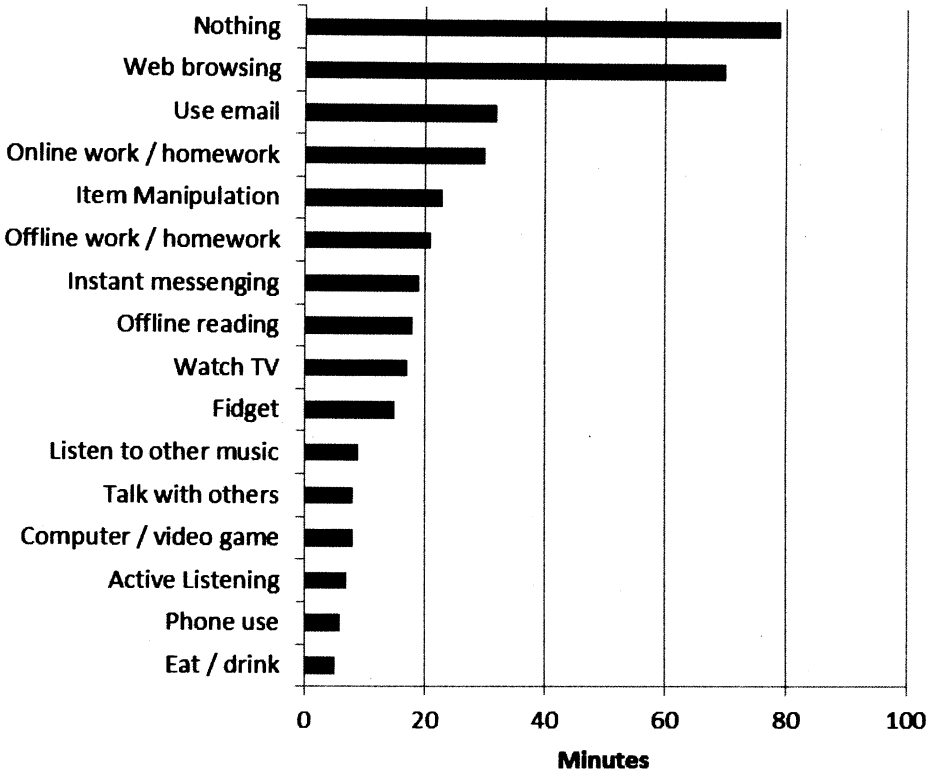


Figure 1- Total task times for recorded behaviors. Computer use for work and leisure is highly represented. While *Doing Nothing* was the single highest task recorded, it only accounts for 22% of the total aggregate on-hold time for the overall sample.

Participants estimated their global on-hold behaviors in the form of percentages, via self-report questionnaire. Figure 2 shows the self-report data as aggregate percentages across all participants. The objective and subjective frequency data are similar in some ways. In both cases, various forms of computer use, including web browsing, e-mail, instant messaging, and academic/professional work are repeatedly the secondary tasks of choice among participants. Furthermore in both datasets the single highest "task" is no task at all, where participants simply waited on hold without engaging in any secondary activities. While the self report data suggest that TV watching, music listening, and

conversation with others in the room are all relatively high frequency tasks, this was not found to be true in the videos.

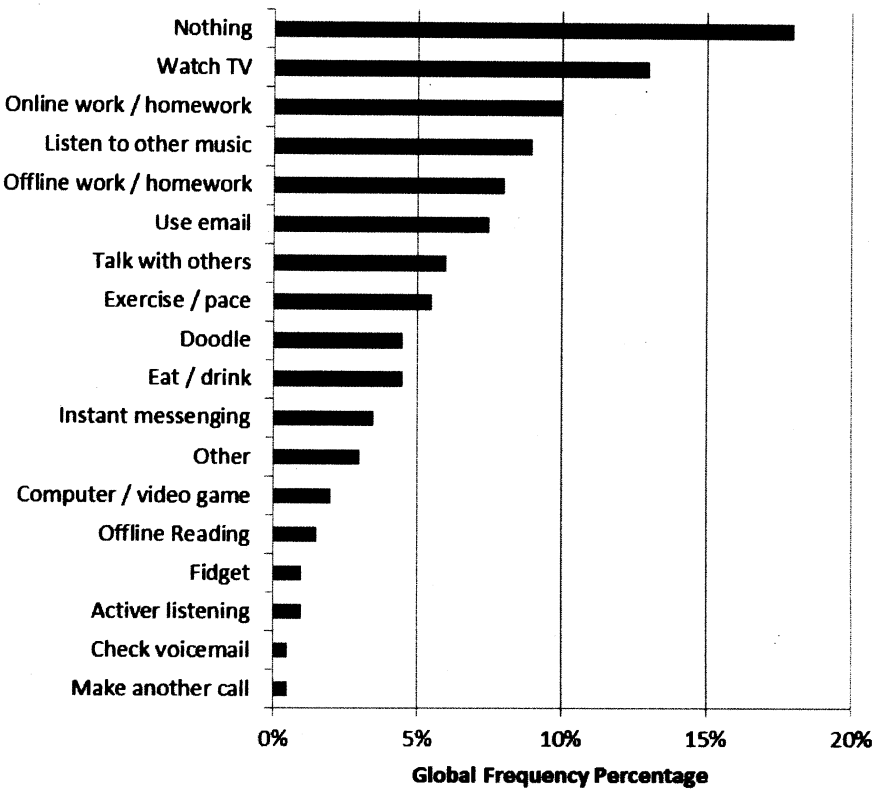


Figure 2 – Aggregated global percentages for self-reported behaviors. As in the video data, the highest single behavior is *Doing Nothing*.

The self-reported secondary behavior data share some characteristics of the observed data, with some key differences. Watching TV and Music Listening are frequently reported but rarely observed. However, Homework and E-mail are both frequently reported and observed.

Table 1- Total Task Time Aggregated Across Participants

Secondary task	Time (Min)	Secondary Task	Time (Min)
Eat/drink	4	Offline reading	18
Phone use	4.5	Instant messaging	19.5
Active Listening	5	Offline work / homework	21
Talk with others	6	Item Manipulation	23
Computer / video game	6	Online work / homework	31
Listen to other music	7	Use email	33
Fidget	12	Web browsing	70
Watch TV	16	Nothing	79.5

DISCUSSION

A number of challenges were encountered during the pilot testing. The naturalistic setting produces variability that is inherent with this type of data collection, and the diverse layouts and lighting conditions of the participants' rooms meant that each trial at a new location had to be planned individually. The camera had a limited field of view of about 90 degrees, and combined with certain room layouts could severely restrict the participants' range of motion, as they were instructed not to stray out of frame. This potentially reduced the number of activities the participants were free to engage in. The presence of the camera and the knowledge of experimental participation also produced behavioral changes in some participants, who remarked during debriefing that they consciously refrained from undertaking secondary tasks. However, even with these limitations, most participants engaged in some sort of secondary activity at least once.

Despite these limitations, the data collected reveals that secondary and tertiary tasks are quite common for on-hold callers, as is task switching. This is especially true once callers knew the process and expected to be placed on hold for minutes at a time. Perhaps due to the population, computer use was the single most frequent task. Computer use is further broken down into casual use (Internet browsing) and more demanding use (homework). Of the tertiary tasks, eating and drinking were

frequent, as participants snacked or sipped while waiting on hold and using a computer. Other tasks included reading and pacing, and some participants spent long periods of time doing nothing except being on hold. These data are consistent with previous self-report data collected by Kortum & Peres (2007).

Both the video and self-report data appear to show that doing nothing is the most frequent single behavior while on-hold. This is likely due to both of the facts that some people are just "holders" who simply do nothing else but wait, and that most people contribute to the "do nothing" category at some point. The "do-nothing" task was operationalized as fifteen seconds or more of inactivity, which is quite common even among busy multi-taskers. Meanwhile, the choice of activity is much more diverse for multi-taskers, whose time is split among several tasks. It is worth noting that "doing nothing" only accounts for 22% of the total aggregate on-hold time for the entire sample, and that the majority of on-hold time across participants was taken up with at least one secondary activity.

Some unexpected behaviors also surfaced. One participant appeared to be actively listening to the hold music, nodding his head in a rhythmic fashion. This is interesting because the mental demands of actively listening to music may be different from that of merely paying enough attention. Consequently, active listening behavior has been differentiated from passive listening in the data. Another unexpected behavior was the physical activity exhibited by many participants. This includes the widest range of behaviors exhibited by the participants, including picking up and putting down objects, making motions with various body parts, interacting with objects in the room such as the blinds, and other such miscellaneous behavior. This behavior was further classified into two types: Fidgeting and Object Manipulation. Object Manipulation involved physical interactions with a clear goal or purpose. For example, one participant was organizing his desk during the call, and sorting various items into bins. On the other hand, fidgeting is a physical interaction without a clear goal, such as twirling a pen. Physical interactions were one of the more commonly observed behaviors in the

video data, and was absent from the self-reported data.

Task selection

The results from the ethnographic study were intended to provide ecologically valid secondary tasks for the workload measurements in Part Two. To that effect, four tasks were chosen to represent highly frequent and diverse behaviors. These were Web Browsing, Math, Reading, and Item Manipulation. These tasks are described in greater detail in Part Two.

Part Two – The effects of APB type and secondary task on mental workload ratings

OVERVIEW

Armed with ethnographically validated representative secondary tasks, the next step was to investigate the relationship between APB type and secondary task, and how they work to influence workload ratings. Since different tasks were expected to have different workload ratings, isolating the effect of APB type from the effect of task type was important. Therefore, the tasks were tested in isolation for workload ratings and performance baselines before being paired with APBs in a multitasking context. Two APB types were investigated - a compositionally constructed electronic stimulus designed to convey a sense of time passage (Stallman, Peres, & Kortum, 2008), and a voice stimulus which periodically updated the caller with queue information.

While previous work in our lab has found an interaction between APB type and secondary task type on workload ratings in dual-task situations, the effect of APB type alone was not found when participants attended to the APBs without the interference of a secondary task (Kortum, Ling, Su, Peres, & Stallman, 2008). We have proposed that this may be due to the construction of previous APB stimuli, which consisted of short sounds lasting a second or less separated by roughly 15 seconds of silence. The high ratio of interstitial silence to stimuli may be contaminating the workload ratings, such

that the ratings are more reflective of the silence than the stimuli. The stimuli in this experiment were constructed to decrease the ratio between silence and stimuli substantially, which was expected to lead to a differentiation between the single task workload ratings of the APBs. This prediction leads to Hypothesis 1.

Hypothesis 1: Overall and component workloads for APBs while multitasking will differ from each other.

Previous research has indicated that the voice prompt is rated higher than silence or tonal APBs in mental workload without multitasking (Kortum et al, 2008). It is hypothesized that this result will be found in a dual-tasking paradigm as well. Combined with Hypothesis 1, this leads to Hypothesis 2.

Hypothesis 2: The voice APB will be rated higher in overall and component workload than the compositional APB while multitasking.

While previous research did not find any significant interaction between the effects of APB type and secondary task, the effect of the large percentage of silence on this interaction is unknown. Furthermore, the new tasks in the proposed studies may conceivably be more mentally demanding than the tasks used in previous studies, which could bring into focus any interaction effects. When the tasks in a dual-task scenario overlap in their input or output modalities, there can often be performance penalties. This phenomenon is well-demonstrated in the literature on multitasking interference (Pashler, 1994). The interference can be caused by response conflict, such as when both tasks require a manual response as opposed to one task requiring a manual response while the other a verbal one (Mcleod, 1977); it can also be the result of a central bottleneck, when both tasks activate overlapping cortical

structures (Klingberg, 1998). Furthermore, there is evidence that dual-task interference may be content-dependent, where each specific pairing of task modalities produces a unique interference effect. Hazeltine, Ruthruff, and Remington (2006) demonstrated that pairing a visual-input vocal-output with an auditory-input manual-output task resulted in twice the cost of the opposite pairing of the same tasks (auditory-input vocal-output and visual-input manual-output). While these performance costs may sometimes be alleviated through practice, they are more apparent for some modalities than others and can rarely be eliminated (Ruthruff, Johnston, & Van Selst, 2001). While specific accounts regarding the underlying mechanisms of modality interference differ, the overall effect is quite robust and has been documented since the first half of the last century. It is reasonable then to propose that the auditory delivery of APBs may produce varying degrees of interference in our dual-task design, depending on the modality of the secondary task.

Hypothesis 3: There will be an interaction effect between APB type and secondary task. APBs will interfere more with secondary tasks of the same modality than tasks of different modalities.

METHOD

Design

The experiment was a completely within-subject design, with all subjects being exposed to all conditions. The independent variables were APB type (2 levels) and task type (4 levels). The dependent variables were NASA-TLX scores and performance measures for each task (see Table 2).

Table 2- Performance measures by task

Task	Web	Object	Math	Reading
Performance Measures	Number of item prices found	Number of completed operations	Number of problem attempted and number of correct solutions	Number of problem attempted and number of correct solutions

Tasks

Four tasks were designed for this experiment, with the results from Part One and prior research serving as the foundation. The tasks were chosen to represent both the objective frequency in the video data, as well as to be diverse and include multiple sensory modalities. These tasks were Math, Reading, Object Manipulation, and Web Browsing. All tasks had a time limit of four minutes, which is the length of the auditory stimuli tested in Part Two.

The Math task consisted of 105 arithmetic problems involving the addition of two random three-digit numbers. The Reading task consisted of 24 multiple-choice questions involving analogies taken from the Reading section of the Graduate Record Examination. These tasks were chosen to reflect the high occurrence of homework in both the self-report and video data, and represented working tasks that callers might engage in while on hold. The length of each task was constructed so that participants could not finish the entire task in time allotted, ensuring that they would be working for the entire duration. For each task, the total number of problems completed and the total number correct were recorded.

The Object Manipulation task consisted of a modified version of the Purdue Pegboard Dexterity Test (PPDT) (Tiffin & Asher, 1948). The PPDT consists of metal pins, collars, and washers, from which the participants make small assemblies by inserting them into holes on a pegboard placed flat on a desk. The modification to the standard PPDT procedure was necessary as the task time for the

baseline test was longer than the 30-second trials outlined by the manual. Participants inserted a pin into a hole, then a collar onto the pin, and finally two washers on top of the collar. Participants could only manipulate a single piece at once, and were thus required to go back and forth between their current position on the board and the component cache, located at the top of the board.

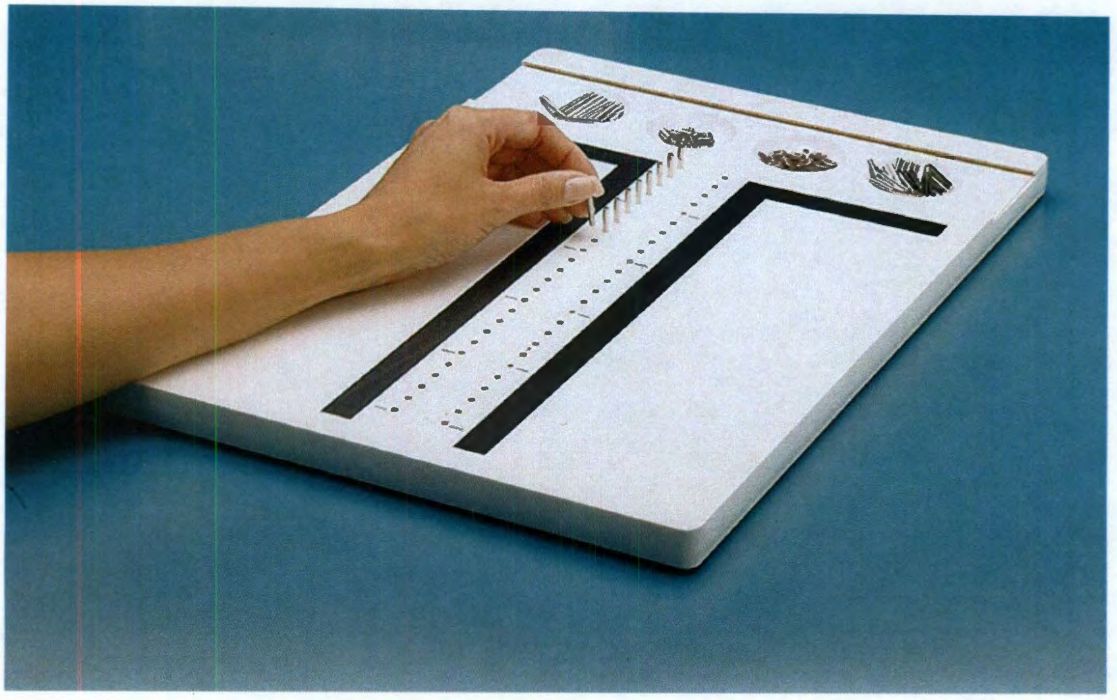


Figure 3 - Purdue Pegboard Dexterity Test. The component cache is at the top of the board, while holes for peg insertion run down the center. Participants were required to interact with only one component at a time, and not grab multiple pieces at once.

Participants were also required to complete a full assembly of pin-collar-washer-washer before moving on to the next hole and assembly. Due to the task time and the number of pieces available in the kit, most participants reached a point where one component was exhausted. They were instructed to dismantle the assemblies they had constructed, one piece at a time, and return the components to their respective caches. If the participant disassembled the entire board, they were to start over and make new assemblies. This continued until the time limit was reached. For each trial, the total number of operations was recorded, where an operation is defined to be the movement of one piece from the cache

to the assembly, or vice versa. All participants used their dominant hand for this task.

The Web Browsing task required the participants to find the price of items on www.amazon.com. The item list consisted of a list of thirteen items, where each item was selected quasi-randomly from a separate category in the drop-down menu on the front page (such as a treadmill from "Sports & Outdoors > Exercise & Fitness"). This task was intended to replicate a web-browsing experience while still providing some metric of performance. Originally the participants were required to find a used version of each item and record that price, but this proved too complex during pilot testing, so the participants were simply asked to find any price for each item. Due to occasional changes in price, participant performance was measured only by number of items found and not price accuracy.

Materials

The main auditory stimuli in this experiment were two APBs, Verbal and Electronic. The Verbal APB consisted of an automated voice that informed the caller of the time left in the hold queue every 10 seconds, by stating: "You have X minutes and Y seconds left on hold". The interstitial time between voice prompts was silent. The voice was a simulated American male, constructed with the ATT Natural Voices Demo. The Electronic APB consisted of a short composition created in collaboration with Dr Kurt Stallman of the Shepherd School of Music. It was a music-like sequence of sounds which conveyed a sense of motion and temporal passage by using a variety of auditory and musical cues. For example, a sense of closure was suggested by chord progressions to signal the end of the stimulus and waiting period, while the addition of instrumentation layers as the stimulus progressed provided a cue of forward motion (Stallman, Peres & Kortum, 2008).

The stimuli were presented via a Javascript program using the Mozilla Firefox web browser. The program had the following functions: maintain cover story, randomize presentation order, put

participants on hold, present auditory stimuli and the NASA-TLX, and record the data. Auditory stimuli were outputted through a standard DTMF telephone handset connected to the PC soundcard, which gave participants the illusion of being on a real phone. As in Experiment II, the Web Browsing task was performed on a Dell Opteron PC with the Mozilla Firefox web browser. The Math and Reading Homework tasks were printed and included in the experimental packet. A stopwatch was used by the experimenter to time the tasks. Three equivalent but different versions of the Web, Math, and Reading tasks were constructed, to provide new stimuli for each of the Baseline, Voice APB, and Electronic APB conditions. No new procedure was constructed for the PPDT because it was thought to be primarily a physical action that would retain its mental demand across trials.

Participants

Participants were 40 students recruited from the Rice University population, screened for normal or corrected-to-normal hearing and vision, and given course credit as incentive. Testing took place in a university laboratory. At the beginning of the experiment the participants were briefed and informed consent was obtained.

Procedure

The testing consisted of three parts (see Figure 4). First participants rated each task for its workload baseline. Next, participants performed the tasks while listening to one of the two APBs. Following a break, the participants performed the tasks again while listening to the other APB. The tasks were presented in random order, and participants had four minutes to work on each task. The experimenter sat behind the participant and kept time with a stopwatch. The Web Browsing task was performed on a Dell Opteron PC with the Mozilla Firefox web browser. The Math and Reading Homework tasks were printed and included in the experimental packet. At the end of each task, the

participants rated their subjective workload during that task with the NASA-TLX.

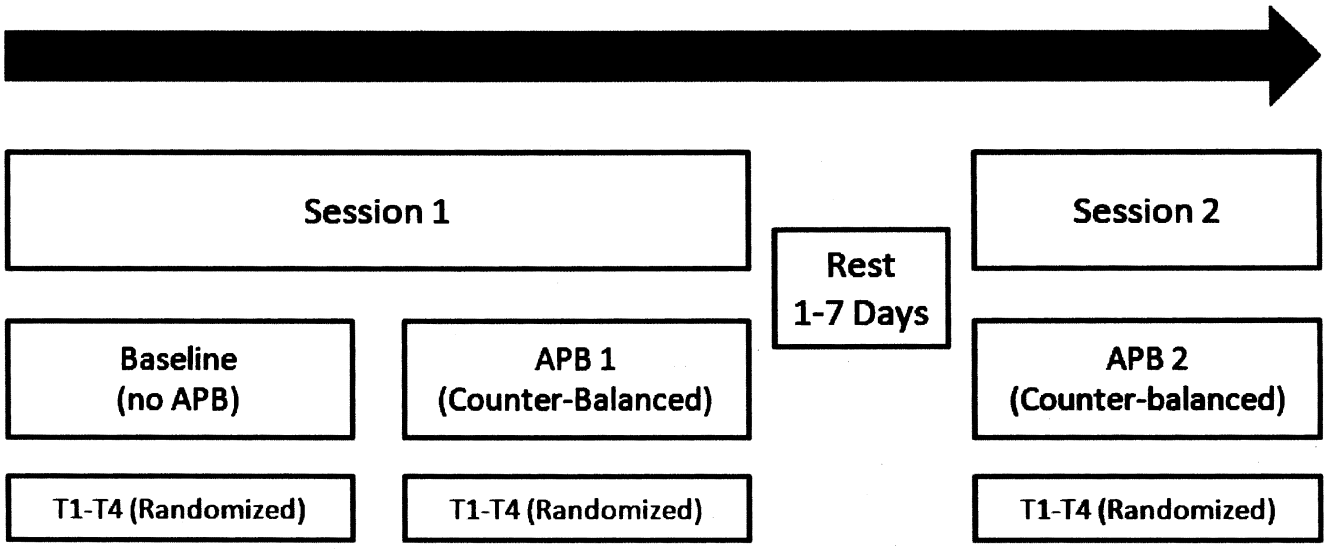


Figure 4- Experimental design. T1-T4 are the experimental tasks. In order to reduce learning and practice effects, the experiment was conducted over two sessions separated by one to seven days.

For the multitasking part of the experiment, participants were told that they were to make several phone calls to a computerized Interactive Voice Response system in order to obtain their account balance. For each call, the participants were placed on hold for 240 seconds before the account balance could be obtained. Participants were notified that they could expect to be placed on hold for some time, but the specific hold length was not revealed. During this hold period, the system played either the Verbal or Electronic APB. Simultaneously, participants performed one of the four secondary tasks. Before each call, the participant was given instructions for the specific task, and asked to work until the call had been answered. At the end of each call, participants obtained the account balance, estimated the amount of time they spent on hold, and rated their subjective workload during the call.

RESULTS

Due to corrupted data, the scores for one participant were excluded from analysis. Overall, the

baseline data indicate that there are differences in workload between the four tasks selected in Part One. The Object task ranked highest in overall subjective workload, while the Web task had the lowest overall workload. Figure 5 illustrates the mean workload ratings for each task in total. A within-subject ANOVA was computed for the effect of task on total workload, which was significant at $F(3,152) = 3.53$, $p = .02$. Bonferroni-corrected post hoc pairwise comparisons did not detect any significant differences between individual tasks.

Next, the NASA-TLX scores were decomposed into their component scales. While the NASA-TLX was designed to output a single global workload rating, its constituent subscales were developed as separate dimensions that provided useful information on their own (Hart & Staveland, 1988, Hart, 2006). A two-way ANOVA found significant effects for task type ($F(3,114) = 8.25$, $p < .01$), scale component ($F(5,190) = 30.17$, $p < .01$), and an interaction between task type and scale component ($F(15,570) = 9.08$, $p < .01$). To decompose the interaction, the simple main effect of task on each of the sub-components of the NASA-TLX was examined by within-subject (see Table 3). The effect of task was found to be significant at the Bonferroni-corrected p-value for the mental, physical, temporal, performance, and frustration components. The effect of task was not significant for the effort component.

Table 3- Effect of task type on NASA-TLX sub-component scores.

	F-value	*p*-value
Mental	$F(3,114) = 16.58$	$p < .001$
Physical	$F(3,114) = 11.97$	$p < .001$
Temporal	$F(3,114) = 4.75$	$p = .004$
Effort	$F(3,114) = 0.81$	$p = .491$
Frustration	$F(3,114) = 4.46$	$p = .005$
Performance	$F(3,114) = 4.68$	$p = .004$

In order to see how the tasks differ on each of the NASA-TLX subcomponents, Bonferroni-corrected pairwise t-tests were performed on each task. Shown in Figure 6, the Object task had significantly higher mental demand than all other tasks. Meanwhile, the Reading task was shown to have higher physical demand than each of the other tasks with (Figure 7), and higher temporal demand than both the Math and Web tasks (Figure 8). Lastly, the Object task was lowest on performance demand (Figure 9), while being highest in frustration (Figure 10). No differences were found on the Effort subscale.

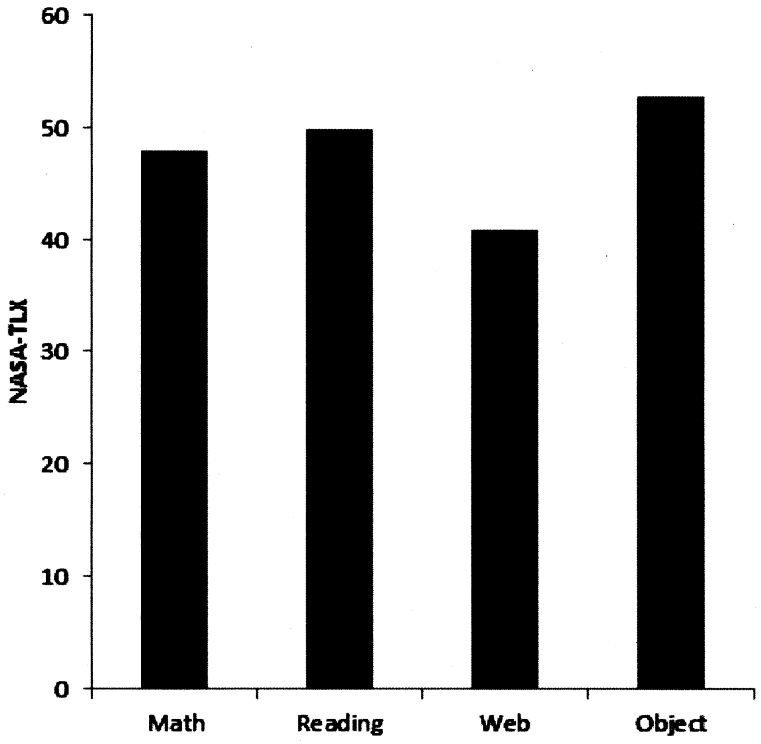


Figure 5- Mean overall workload ratings by task.

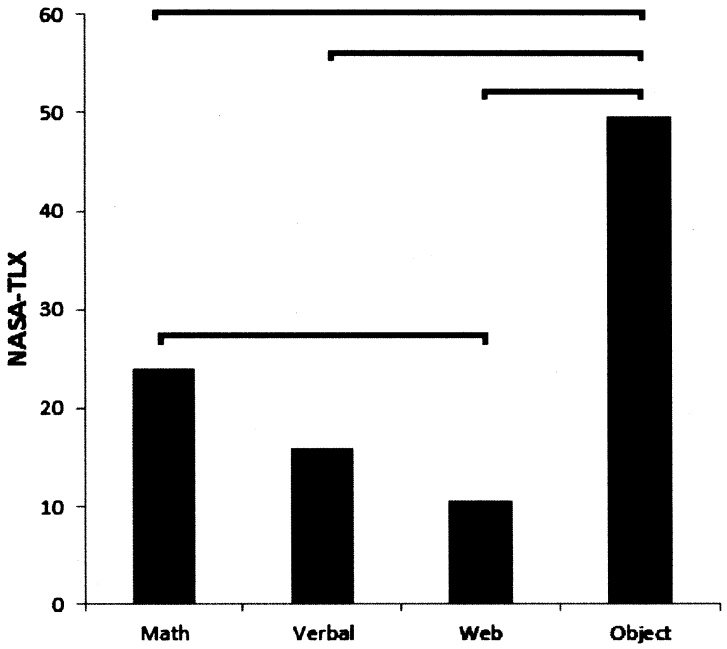


Figure 6- Mean mental workload ratings by task. Bracketed bars indicate a Bonferroni-corrected significant t-test.

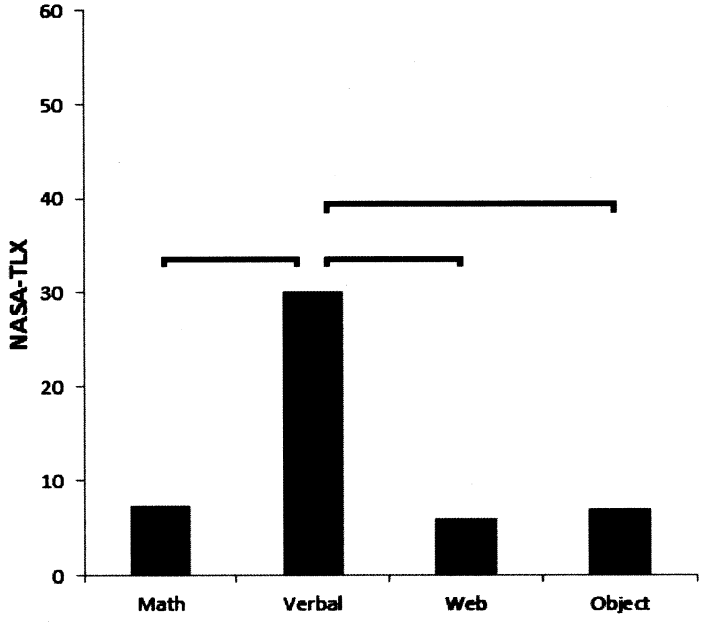


Figure 7- Mean physical workload ratings by task. Bracketed bars indicate a Bonferroni-corrected significant t-test.

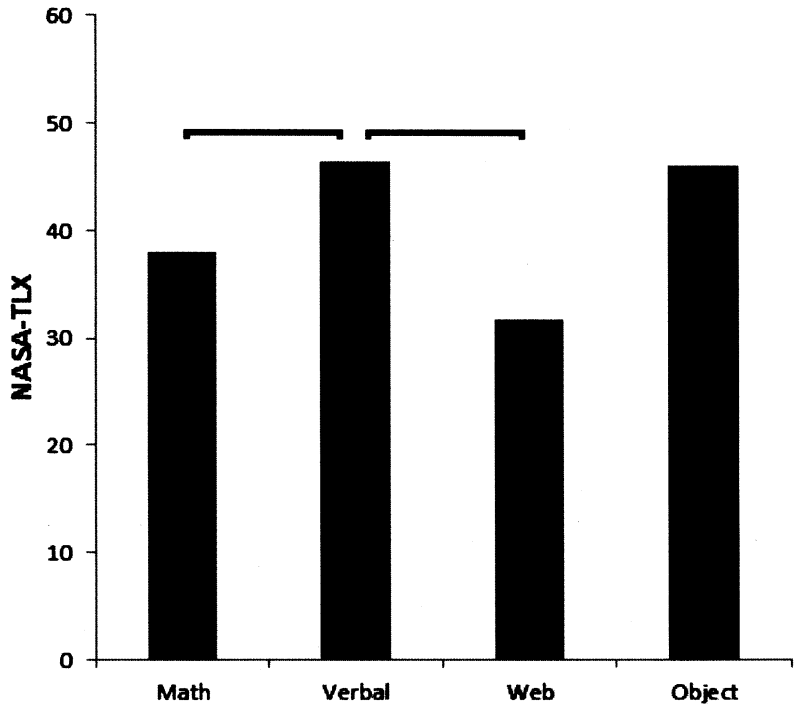


Figure 8- Mean temporal workload ratings by task. Bracketed bars indicate a Bonferroni-corrected significant t-test.

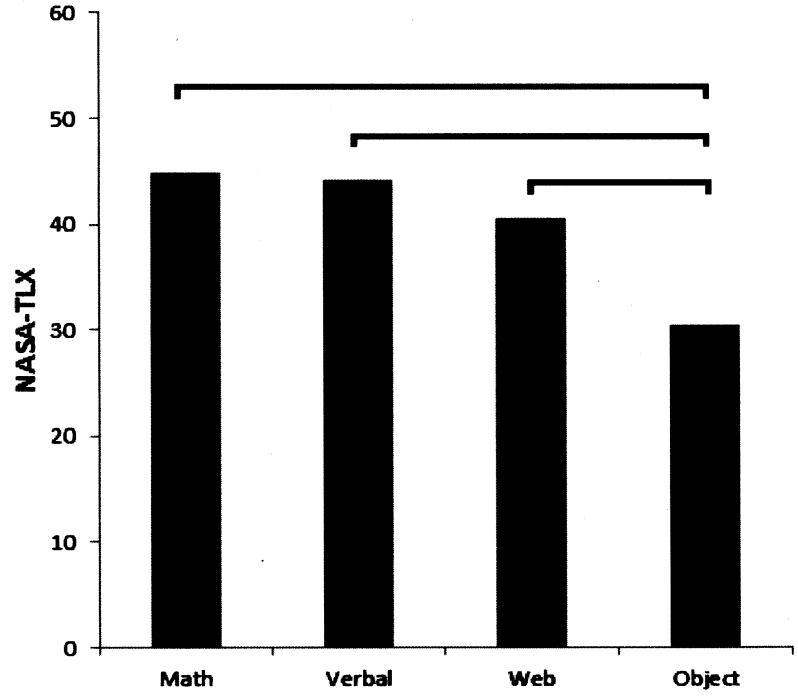


Figure 9- Mean performance demand by task. Bracketed bars indicate a Bonferroni-corrected significant t-test.

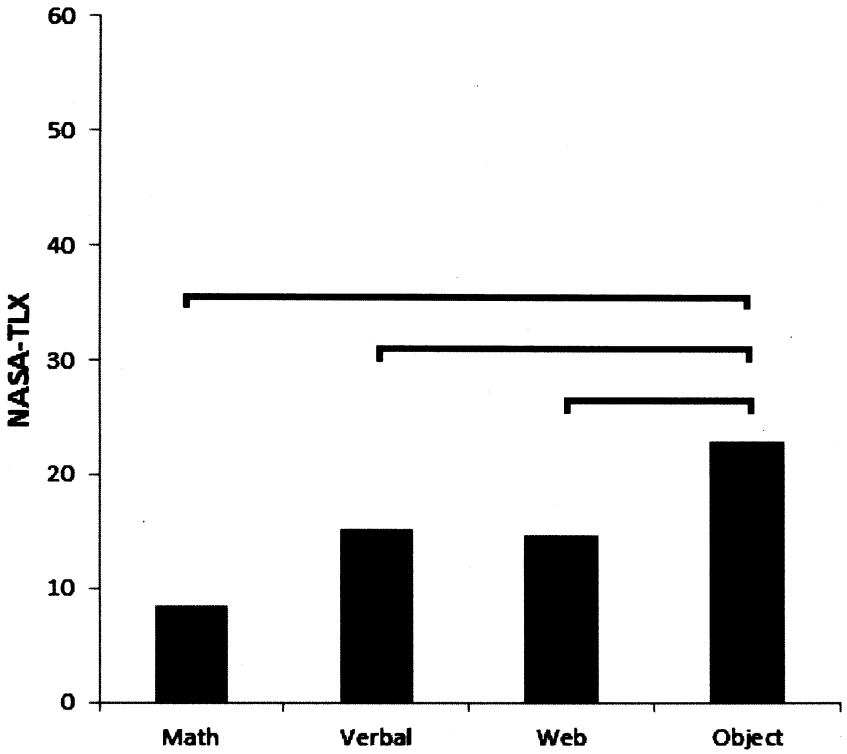


Figure 10- Mean temporal workload ratings by task. Bracketed bars indicate a Bonferroni-corrected significant t-test.

Correlation matrices among the decomposed subscales were computed for each task. Overall, correlations were low for these data. No subscales were found to correlate significantly with each other for the math task (see Table 4). For the object task, the Mental subscale correlated negatively with the Physical subscale, $p < .01$, the Physical subscale correlated with the Frustration subscale, $p = .02$ and the Performance subscale correlated negatively with the Frustration subscale, $p = .03$ (see Table 5). For the reading task, the Mental and Physical subscales were negatively correlated, $p < .01$, and the Performance subscale was negatively correlated with the Frustration subscale, $p = .01$ (see Table 6). Lastly, on the web task, the Mental subscale was correlated with the Effort subscale, $p = .01$. (see Table 7).

Table 4- Correlation matrix for math task.

		1	2	3	4	5	6
1	Mental	1.00	-.29	.16	-.27	.13	.05
2	Physical		1.00	.02	-.29	.10	-.19
3	Temporal			1.00	-.01	.09	-.12
4	Performance				1.00	.08	-.10
5	Effort					1.00	.11
6	Frustration						1.00

Table 5- Correlation matrix for object task.

		1	2	3	4	5	6
1	Mental	1.00	-.57**	-.09	-.18	.26	.02
2	Physical		1.00	-.03	.28	-.27	.39*
3	Temporal			1.00	.03	-.01	-.34*
4	Performance				1.00	-.08	-.25
5	Effort					1.00	.06
6	Frustration						1.00

Note: * $p < .05$. ** $p < .01$

Table 6- Correlation matrix for reading task.

		1	2	3	4	5	6
1	Mental	1.00	-.55**	.26	.09	.09	-.19
2	Physical		1.00	-.05	-.19	.23	.29
3	Temporal			1.00	-.25	-.02	-.05
4	Performance				1.00	-.07	-.40*
5	Effort					1.00	.09
6	Frustration						1.00

Note: * $p < .05$. ** $p < .01$

Table 7- Correlation matrix for web task.

		1	2	3	4	5	6
1	Mental	1.00	-.07	-.08	-.03	.39*	.26
2	Physical		1.00	-.13	-.20	.14	-.02
3	Temporal			1.00	-.11	-.04	.20
4	Performance				1.00	.21	-.21
5	Effort					1.00	.21
6	Frustration						1.00

Note: * $p < .05$

The results from the multitasking sessions indicate that for overall workload, the effect of APBs on workload ratings is different for the various tasks. A repeated measures ANOVA on overall NASA-TLX scores results in a significant interaction between the effect of APB and the effect of task, $F(3,111) = 3.45, p = .02$. Neither the main effect of APBs nor that of task was statistically significant.

To further examine the variation in difficulty between the tasks themselves, difference scores were created from the multitasking workload scores and the baselines (see Figure 11). The graph clearly illustrates that the increase in workload for the math task combined with the electronic APB is higher than other Task/APB combinations.

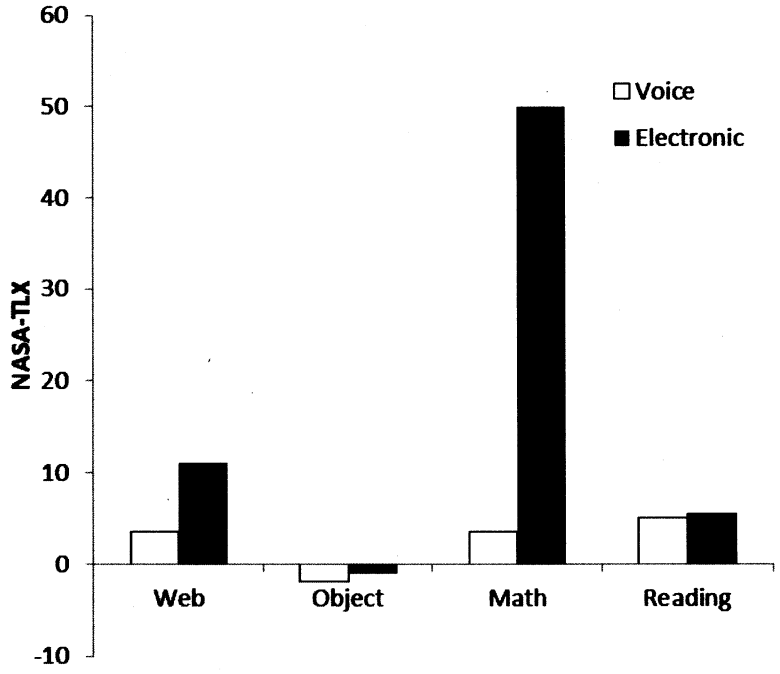


Figure 11- Overall workload difference scores, comparing against baseline task workload.

An ANOVA on the difference scores detected a significant interaction between APB type and task, $F(3,111) = 52.8, p < .01$. Additionally, each of the main effects of Task ($F(3,111) = 46.8, p < .01$) and APB type ($F(1,37) = 22.7, p < .01$) was also significant. From the plot in Figure 11, it is apparent

that this interaction is being driven by the difference between the voice and electronic APBs on the math task. Post hoc comparisons of the two APBs at each of the tasks confirm this hypothesis, with the only significant test being that of the math task, $F(1,37) = 148.9$, $p < .01$.

As in the baseline analysis, the NASA-TLX scores were then decomposed into subcomponents for more detailed analysis. A three-way repeated measures ANOVA was used to investigate the relationship between APB type, task type, and subcomponent workload ratings. The APB variable had three levels – baseline (no APB), voice, and electronic. The three-way interaction was significant, $F(30.1080) = 4.45$, $p < .01$. All three two-way interactions were also significant, as well as all simple main effects, except for the effect of APB type (see Table 8).

Table 8- Repeated measures ANOVA results for three-way analysis.

Effect	*F*-value	*p*-value
APB*Task*Component	$F(30, 1080) = 4.45$	$p < .01$
Task * Component	$F(15,540) = 10.04$	$p < .01$
APB*Component	$F(10,360) = 7.50$	$p < .01$
APB*Task	$F(6,216) = 10.74$	$p < .01$
Component	$F(5,180) = 27.55$	$p < .01$
Task	$F(3,108) = 17.96$	$p < .01$
APB	$F(2,72) = 1.62$	$p = .21$

To decompose the three-way interaction, the data were collapsed on the task variable and the APB*component interaction was examined for each task. For the web task, the APB*component interaction was significant, $F(10,360) = 2.54$, $p = .02$ (Greenhouse-Geisser corrected, see Figure 12). Bonferroni-corrected post hoc comparisons for the main effect of APB on each subcomponent scale failed to achieve significance.

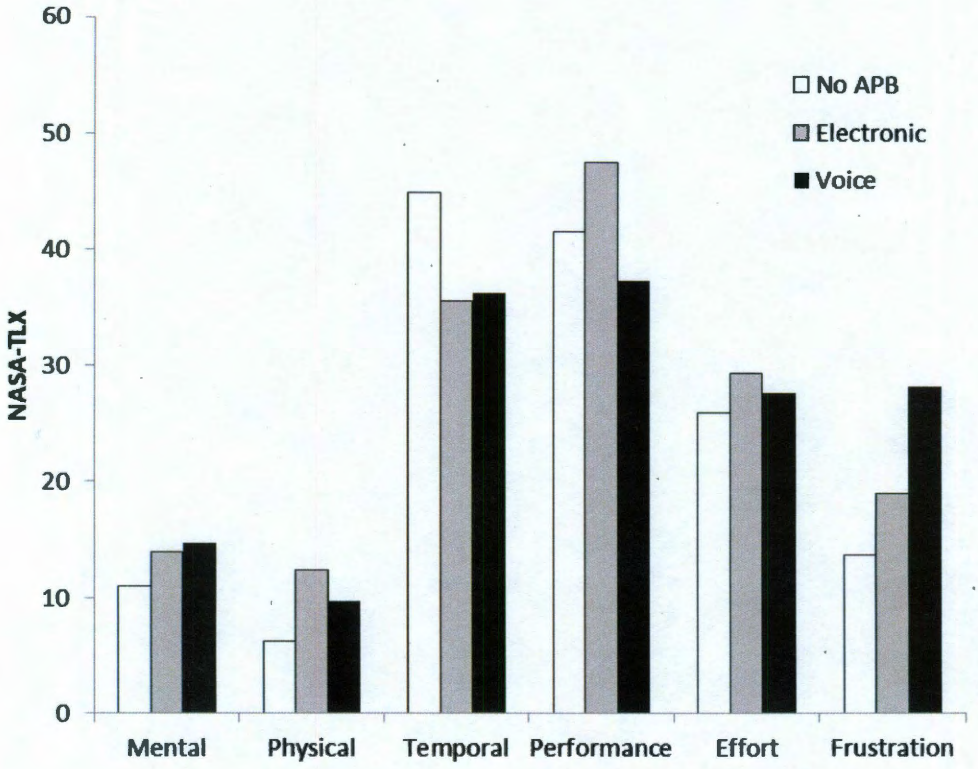


Figure 12- Decomposed workload data for the web browsing task.

For the object task, the APB*component interaction was significant, $F(10,360) = 4.98$, $p < .01$ (see Figure 13). The main effects of component and APB were both significant, with $F(5,180) = 16.20$, $p < .01$ and $F(2,72) = 4.14$, $p = .02$, respectively. The tests of simple main effect of APB for the mental component was significant, $F(2,72) = 10.89$, $p < .01$. Pairwise comparisons among levels of APB on the mental component revealed that both APBs had lower mental workload than the baseline (no APB) condition, with $F(1,36) = 5.66$ and $p = .01$ for the electronic APB and $F(1,36) = 18.49$ and $p < .01$ for the voice APB. For the temporal component, the simple main effect of APB did not make the Bonferroni-corrected significance criterion, but the pairwise comparisons did indicate that the baseline condition had higher temporal demand than the electronic APB, $F(1,36) = 8.55$, $p < .01$.

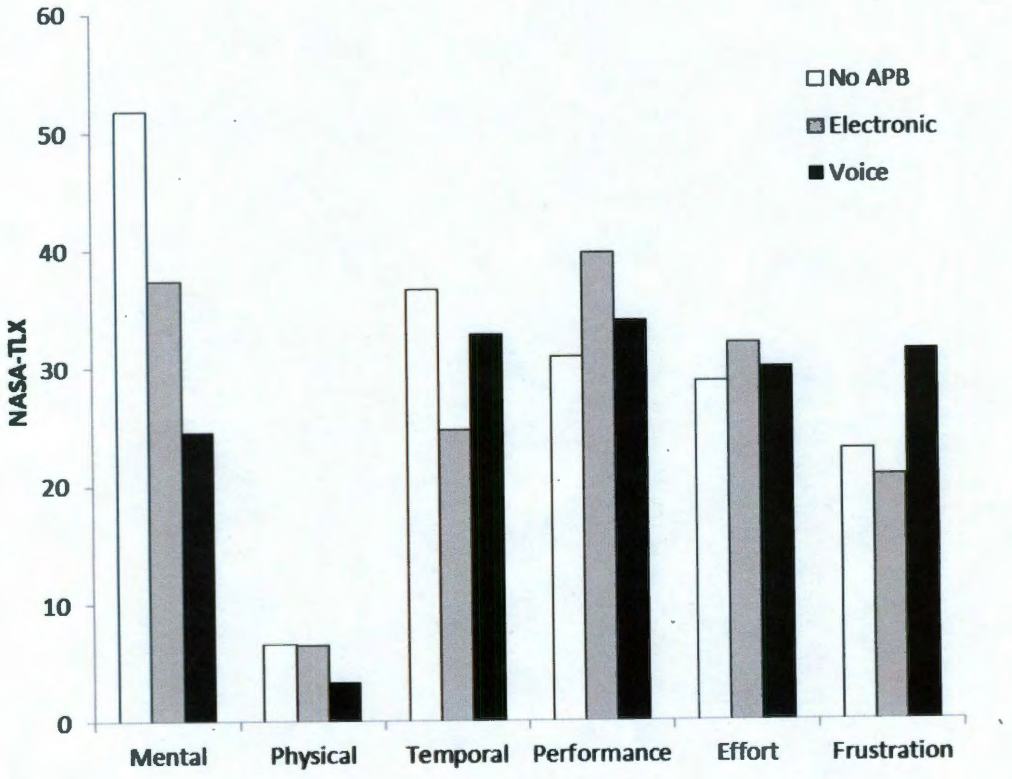


Figure 13- Decomposed workload data for the object task.

For the math task, the APB*component interaction was significant, $F(10,360) = 3.03$, $p < .01$ (see Figure 14). The main effect of component was also significant across all components, $F(5,180) = 26.28$, $p < .01$. For the temporal component, the main effect of APB was significant, $F(2,72) = 7.39$, $p < .01$, and the baseline condition has higher temporal demand than the electronic APB ($F(1,36) = 13.84$, $p < .01$) and the voice APB ($F(1,36 = 7.05$, $p = .01$).

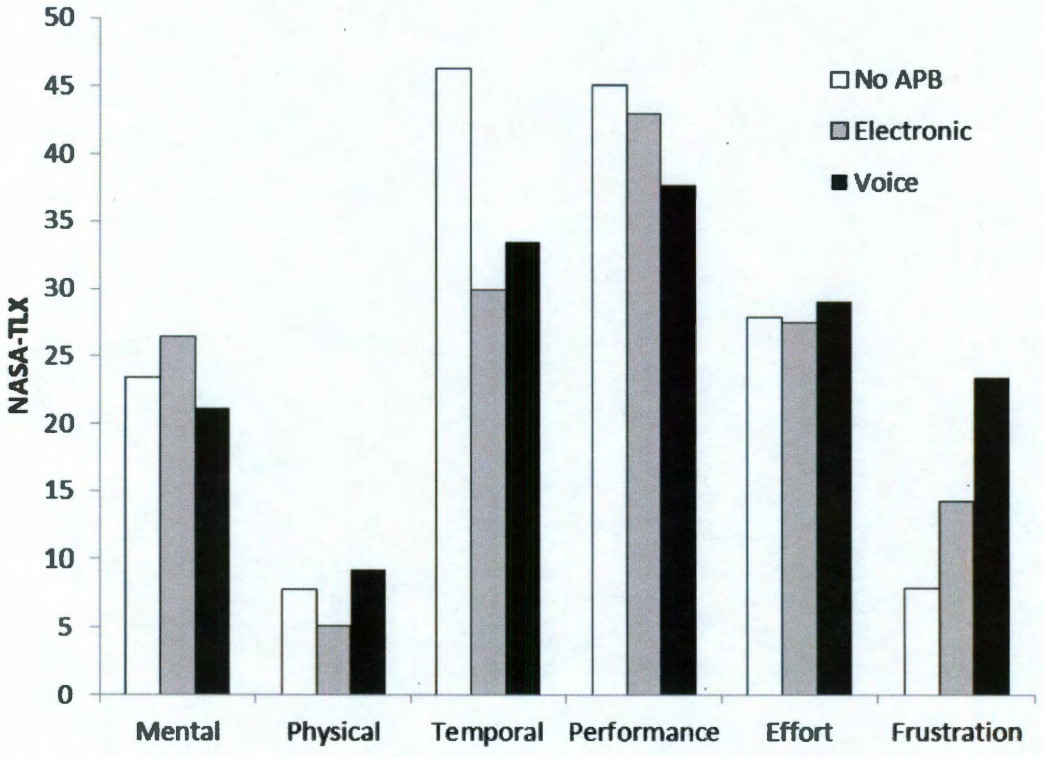


Figure 14- Decomposed workload data for the math task.

For the reading task, the APB*component interaction was significant, $F(10,360) = 10.85$, $p < .01$ (see Figure 15). The main effect of component was significant, $F(5,180) = 10.24$, $p < .01$, as well as the main effect of APB across all levels of component $F(2,72) = 15.84$, $p < .01$. The effect of APB was significant on the mental subcomponent, $F(2,72) = 50.01$, $p < .01$, and the electronic APB had higher mental demand than both the baseline condition ($F(1,36) = 85.78$, $p < .01$) and the voice APB ($F(1,36) = 87.04$, $p < .01$).

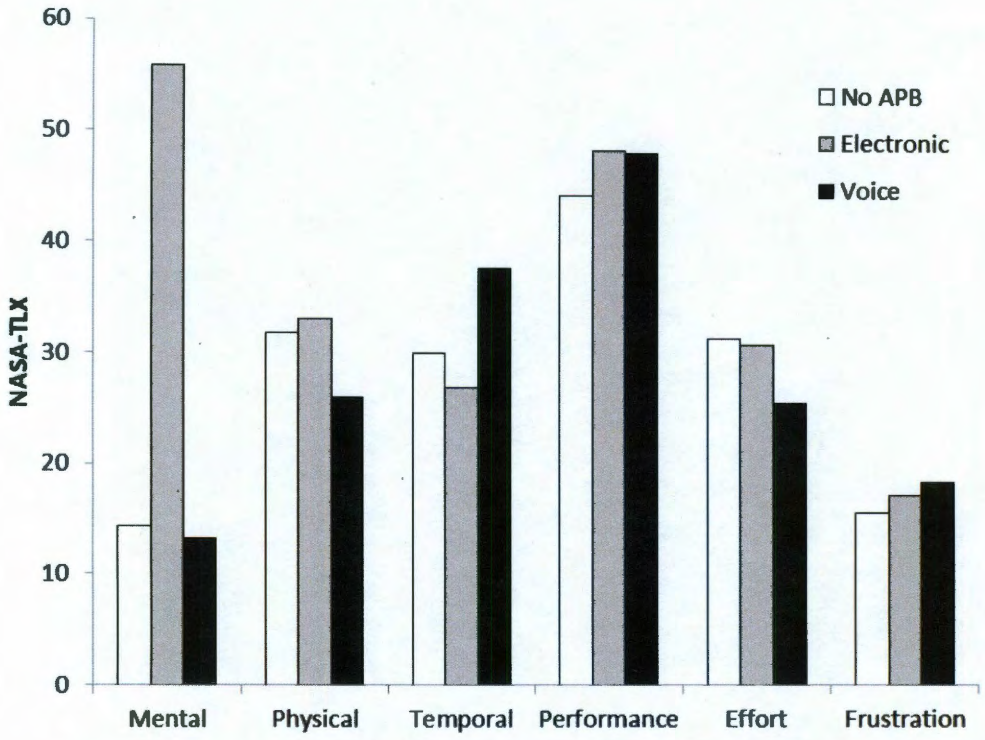


Figure 15- Decomposed workload scores for the reading task.

Apart from the workload data, objective performance data was also recorded in all sessions as an additional measure of workload. Since the tasks were selected to be different from one another, their performance measures naturally diverge and are not well suited to direct comparison. For the web browsing task, the performance metric was number of items completed (see Figure 16). Accuracy was not measured due to price changes by the retailers. Repeated measures ANOVA did not find a significant effect of APB type, but the effect of session was significant, $F(2,72) = 8.71, p < .01$. Trend analysis showed a significant linear effect ($F(1,36) = 11.53$) as well as a significant quadratic effect ($F(1,36) = 37.30, p < .01$.

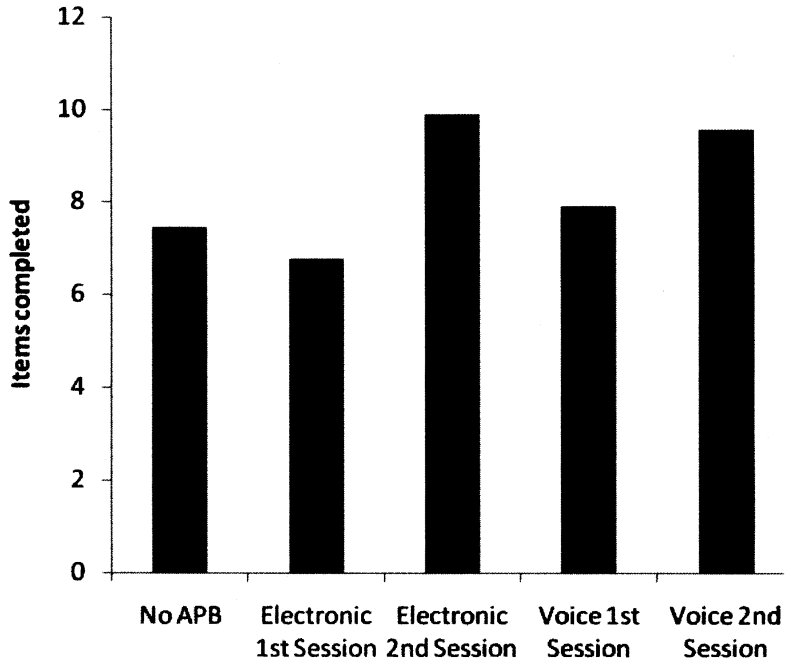


Figure 16- Performance data on the web task, broken down by session.

Similarly, the object task was measured on the number of operations completed before the time limit (see Figure 17). Repeated measures ANOVA revealed an effect of APB type, $F(2,78) = 6.84, p < .01$. However, no significant effect of APB type or session was found. The effect of APB type on Manual task performance seemed to be driven by the difference between the baseline condition and the

two experimental conditions. A t-test of the two APB types alone was non-significant.

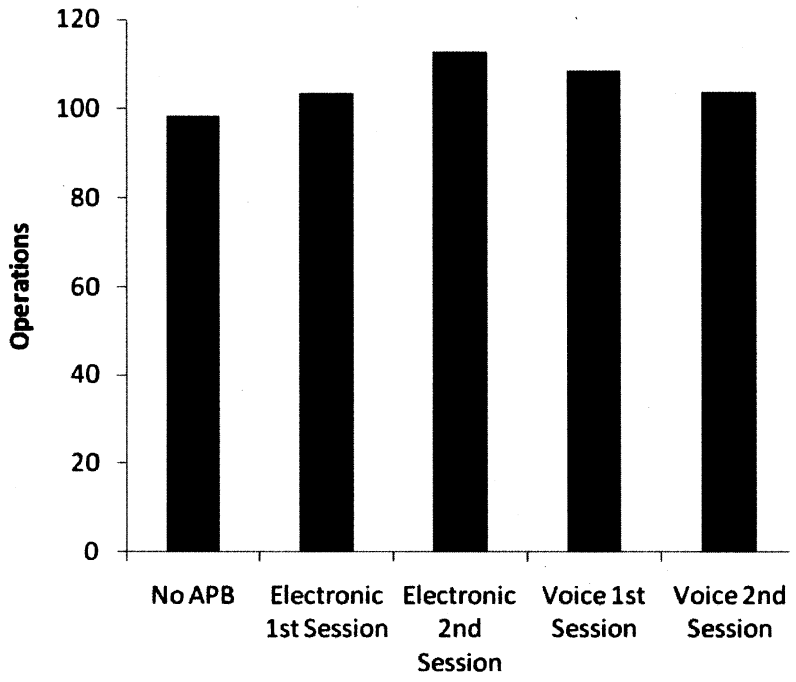


Figure 17- Effect of APB type on object task performance, broken down by session.

For the math and reading tasks, there were more performance measures. For each task, the total number attempted and the total number of correct responses was recorded. For the reading task, an ANOVA of the effect of APB type on total number of problems attempted was significant, $F(2,78) = 13.53$, $p < .01$ (see Figure 18). The effect of session was also significant, $F(2,72) = 5.09$, $p < .01$, with a quadratic trend, $F(1,36) = 8.04$, $p < .01$. For the number of correct responses, The effect of APB type on number of correct answers given was significant, $F(2,78) = 5.80$, $p < .01$ (see Figure 19). The effect of session was significant as well, $F(2,72) = 4.00$, $p = .02$, and a quadratic trend was found, $F(1,36) = 6.48$, $p = .02$. On the other hand, no statistically reliable effects were found for the Math task performance (see Figures 20 and 21).

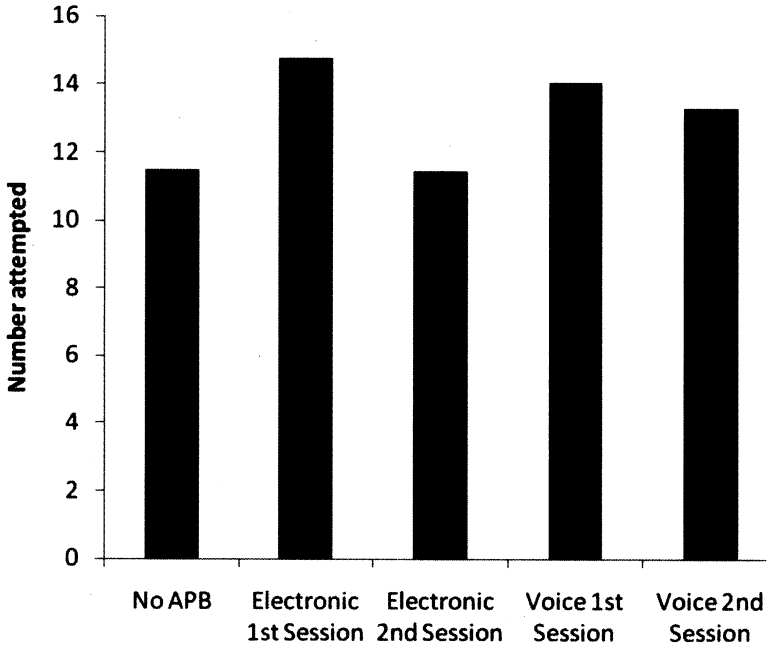


Figure 18- Total number of items attempted for the reading task.

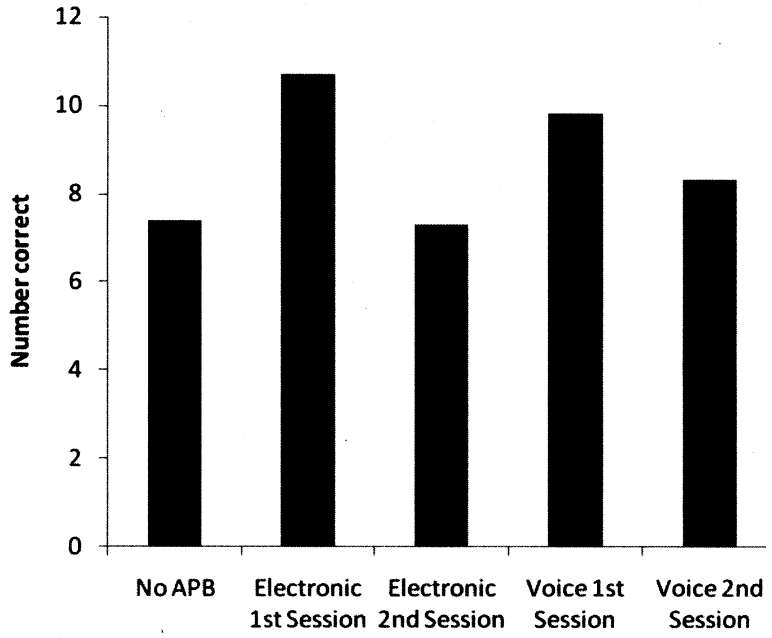


Figure 19- Number of correct responses for the reading task.

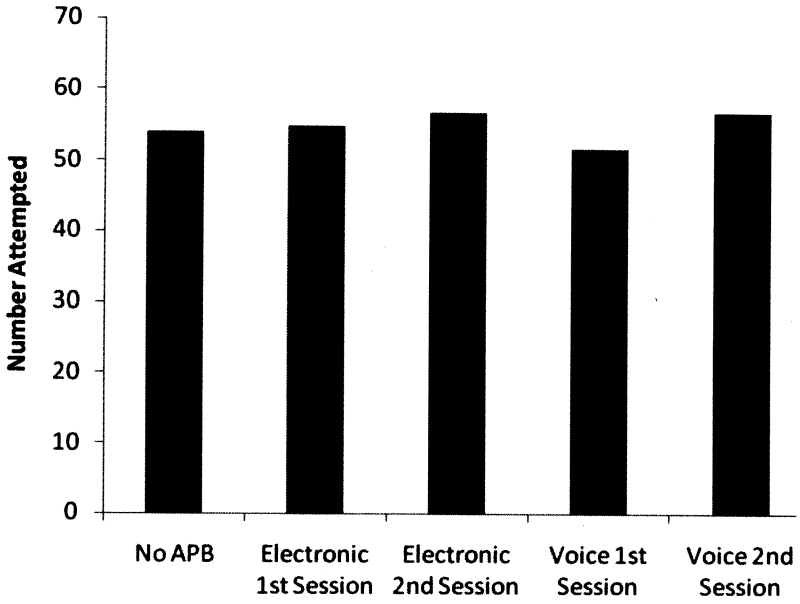


Figure 20- Total items attempted for the math task.

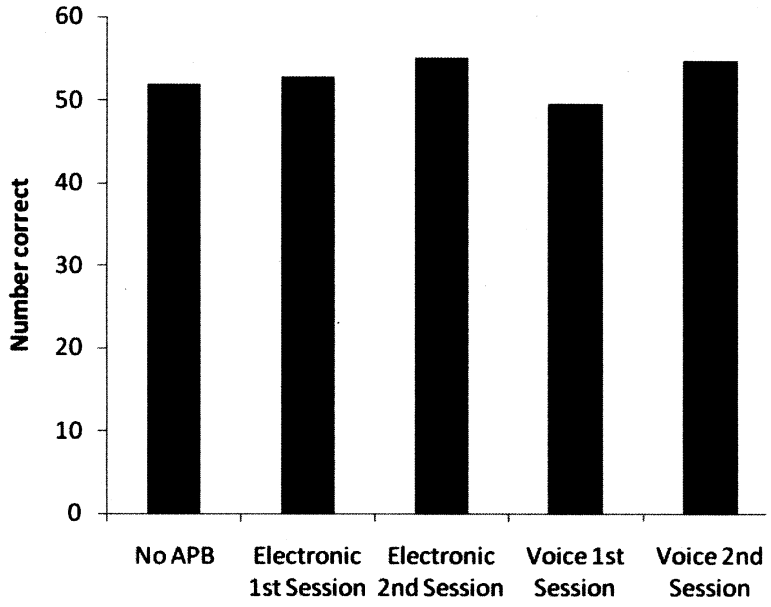


Figure 21- Total number correct for the math task.

To dovetail with previous research on APBs, participants' estimation of their hold time was also recorded. Since hold time was not manipulated for this experiment, all trials had a time limit of 240 seconds. A difference score was computed from the participants' estimates of their hold time. A repeated-measures ANOVA found a significant main effect of APB type on time estimation, $F(1,37) = 6.15$, $p = .02$. The APB*task interaction was not significant. The electronic APB resulted in underestimation, while the voice APB resulted in overestimation, though this overestimation was not significantly different from zero (see Figure 22). The absolute magnitude of the estimation error was greater for the electronic APB ($M = 51.28$, $SD = 45.77$) than the voice APB ($M = 16.75$, $SD = 29.02$), $t(37) = 4.22$, $p < .01$ (see Figure 23).

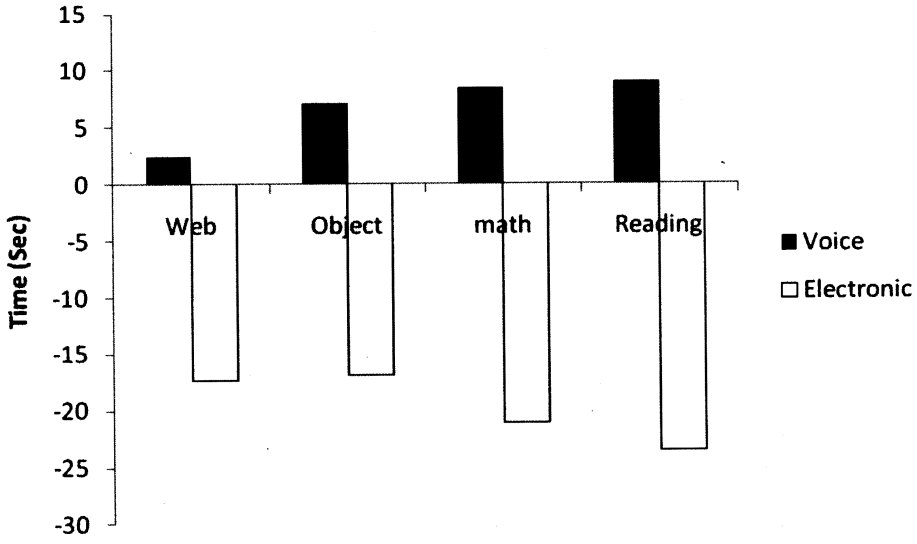


Figure 22- The effect of APB type and task on estimates of hold time.

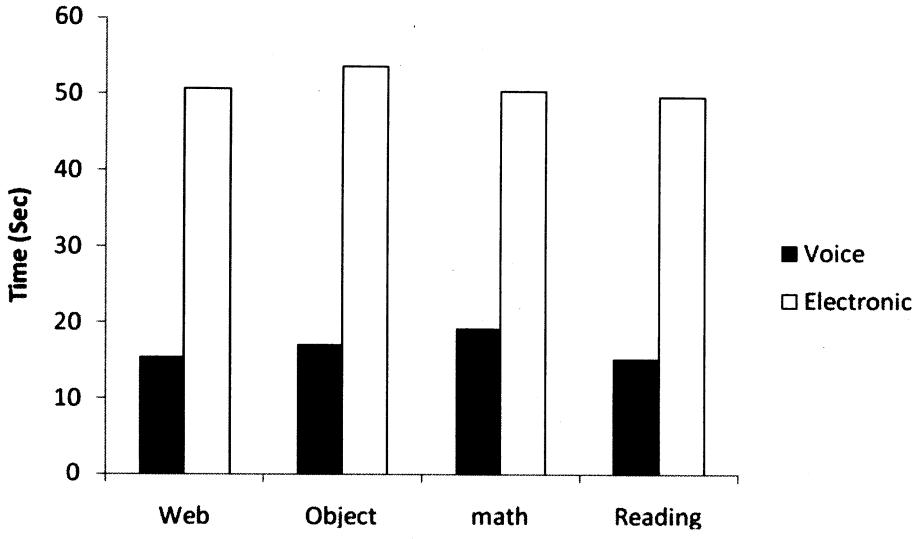


Figure 23- Absolute hold time estimation errors by task and APB.

DISCUSSION

A particularly interesting set of results from the present study involves the predictions made by Multiple Resource Theory (Wickens, 2002). According to the four-dimensional stage model proposed by Wickens, simultaneous tasks that occupy different positions on one or more dimensions in the Multiple Resource Model should exhibit performance or workload savings, while those tasks that occupy overlapping segments of the dimensions will exhibit some costs. Our multitasking procedure was intended to provide some amount of conflict amongst the disparate resources, but that conflict was not consistently observed. For example, on the dimension of perception / cognition versus response, both the primary listening and secondary ethnographic tasks can be argued to have a large perceptual/cognitive component. Listening and attending to queue information and solving reading comprehension problems both require perception and cognition, yet very little performance decrement was seen. On the other hand, the PPDT can be argued to have a larger footprint on the response side of the scale, requiring acute spatial awareness and proprioperception, which should produce performance savings when paired with principally perceptive tasks like APB listening. However, participants rated the PPDT to be the most demanding of all the tasks. Another example is the processing code dimension, consisting of spatial and verbal coding. Tasks such as web browsing and reading would seem to invoke verbal faculties more than a task like the PPDT, and thus be subject to more interference from a verbally-based APB. Again, this was not the case for our data. Wickens does provide that the Multiple Resource Model performs better for high demand tasks than low demand ones, and the tasks and stimuli used in the study are of low to moderate demand when compared with other tasks that have been measured with the NASA-TLX (Young, Reilley, Grasha, Bishop, Lis, and Roberts, 2000, Schmutz, Heinz, Metrailler, and Opwis, 2009). While a detailed treatment of the Multiple Resources Model and its theoretical boundaries is beyond the scope and intent of this paper, it

does provide a very useful framework for viewing and interpreting some of the unusual data patterns.

In terms of overall task demand, the Purdue Pegboard Dexterity Test appeared to be the highest. This is somewhat surprising, as the PPDT is primarily a physical task which requires little thought once the procedure has been learned. However, compared with the other tasks, the PPDT is the least familiar and most novel for the participant population. This novelty may have contributed to the high workload ratings, due to participants having to learn an entirely unfamiliar task and internalize instructions. In addition, the high ratings of frustration and temporal demand probably also played a role. By comparison, the web task scored low across the board from overall workload to the sub-component scales, which fits well with the behavioral data gathered in Part One. That is, it is reasonable for participants who engage in a large amount of Internet activity to be comfortable with it, and consequently give low subjective workload ratings.

The high physical rating for the reading task is a befuddling result with no clear explanation, other than noisy data as evidenced by a large variance. While the physical demand ratings for the other tasks are mostly quite low, the distribution for the reading task is clearly much higher (See Figure 24). After repeated checking for coding errors and administrative faults, there was no indication that these data in any way misrepresent the responses of the participants. The reading task was constructed to be one of the more difficult tasks in the experiment. One possibility is that some participants may have felt a physical arousal from the challenge, particularly under time pressure. Another is that participants misunderstood 'Physical Demand' as being overall demand, despite the fact that the subscales were explained to them, in addition to detailed descriptions of the sub-scales being immediately available. This points to a potential limitation of the NASA-TLX, in that the questionnaire is subject to participant interpretation even after careful instruction, and may not be measuring what it is expected to be measuring. Beyond the NASA-TLX, this is a challenge to the validity of subjective workload measurement in general, and a strong argument for augmenting subjective measures with other data

sources.

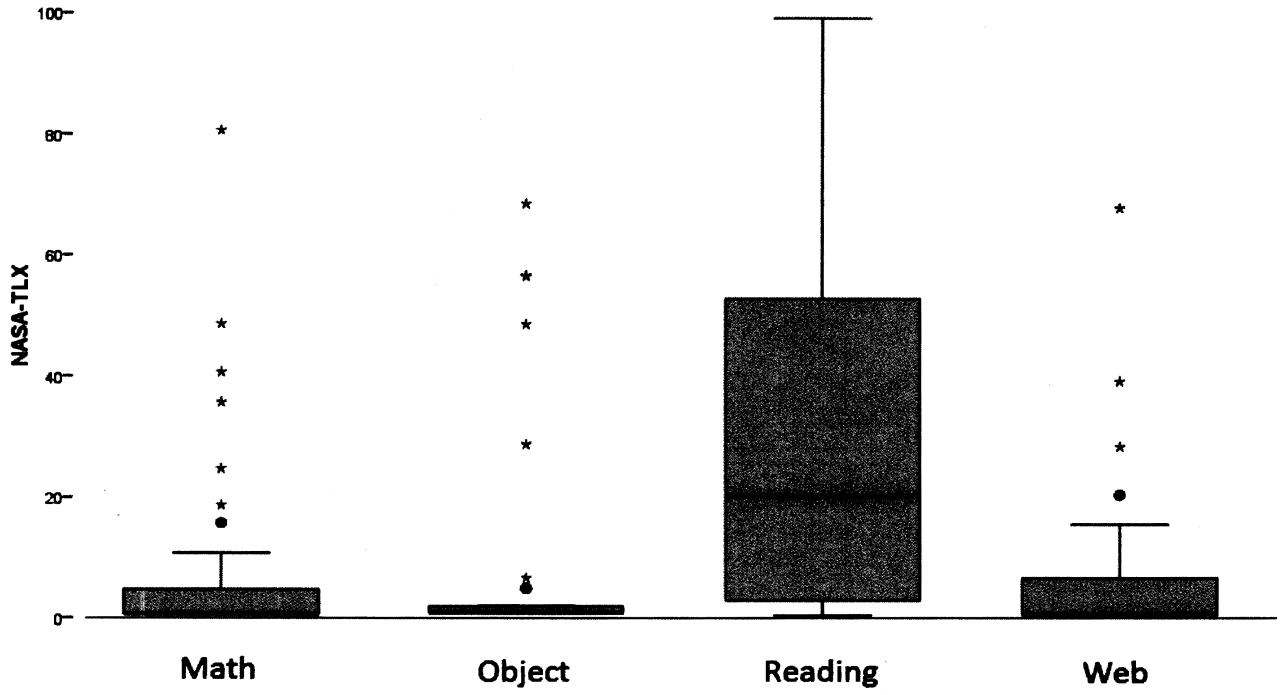


Figure 24 – Distributions for the ratings on physical demand, by task. Participants rated reading to be much more physically demanding than other tasks.

The baseline data demonstrate that the four tasks selected from the behavioral and self-report data do have varying levels of both overall workload and each of the sub-components of the NASA-TLX. This validates the importance of establishing baselines in this series of experiments, as the inherent workload differences in the tasks would have made it harder to observe the effects of APBs.

Many of the results from the experiment were unanticipated. However, they are nevertheless interesting and implicative of a number of conclusions. In relation to the predictions made in the outlining stages of the paper, we have found some support for the hypotheses, though not always in the direction anticipated. A major theme is that for these tasks, the effect of APBs on workload is quite small and easily overwhelmed by other influences.

Predictions

Hypothesis 1: Workloads for APBs while multitasking will differ from each other.

While it is difficult to interpret clearly, the APB*task*component three-way interaction does suggest that APB type plays a part in determining subjective workload for the participants. There are clearly differences between APB types, especially on the mental demand workload component for the reading and object tasks (see Figures 13 and 15). Web browsing as performed in these experiments is essentially a recognition activity, where participants simply had to determine if the product on the screen was the correct one. Compared with performing arithmetic and analogies, web browsing may be less demanding, which is supported by the differences in these tasks' baseline workloads. However, the object task performed in the opposite direction of expectations, dominating both the overall TLX score and even the mental demand component. At first glance, it should appear that the object task should require little cognition compared with the other tasks. One explanation may be the novelty of the Purdue Pegboard Test to the participant population, which has long since become accustomed to web surfing and homework-like tasks. The object task also involved the most amount of instruction, both due to its novel nature and the limitations of the equipment. Furthermore, execution of the object task proceeded at a more rapid rate than the other tasks, with each operation only taking a couple of seconds. It is unclear what the equivalent operator units are in the other tasks, as they differ depending on the granularity of analysis. For example, an operator unit in the web browsing task might be a single keystroke, or the completion of an entire item response. These factors may have led, at least in part, to differences in the temporal demand component on the task ratings. Another reason for the unusual results may simply be lack of demand. None of the tasks were especially difficult, since they reflected what real callers can typically do with little effort during real calling conditions. The low-demand tasks may have not been sufficient to exhaust the mental resources of the participants, thus producing what is

essentially noisy data as opposed to a clear pattern in the expected direction. Indeed, a significant practice effect was found for the performance data, with a session effect in three of the six performance measures, and the APB effect largely being driven by the differences between the baseline condition and the multitasking conditions. The fact that the practice effect is manifest over and above the effect of APB type may be indicative of the small effect and large noise.

When comparing the multitasking conditions with the baseline task workloads, a strong difference between the two APBs emerges, but only on the math task. Both the interaction effect and the main effect of APB type support the hypothesis, with the caveat that it may not apply to all secondary tasks. The significant effect of APB type on hold time estimation is further evidence in support of the hypothesis that these two APBs are in fact different in subjective workload in a multitasking situation.

Hypothesis 2: The voice APB will be rated higher in workload than the compositional APB while multitasking.

The data form the opposite pattern than what is predicted. For overall workload, the electronic APB was found to have higher multitasking – baseline difference scores than the voice APB on the math task. Furthermore, on the mental demand component for the reading task, the compositional electronic APB was rated higher than the voice APB. Otherwise, in no case was the voice APB found to be significantly more demanding than the electronic APB. Confusingly, the baseline (no APB) condition was found to have higher demand than one or both APBs in several cases (mental and temporal demand for the object task, and temporal demand for the math task). As found in previous research, stimuli with long interstitial silences may have deflated workload ratings due to the silence portion of the stimuli dominating the rating. Even though the Voice APB in this experiment was

designed with a shortened ISI, it was compared against an Electronic APB which had no ISI at all. Meanwhile, the objective performance data does not point to either APB interfering more than the other. Rather, the performance data is dominated by a marked practice effect, and does not differentiate between the two types of stimuli.

Hypothesis 3: There will be an interaction effect between APB type and secondary task. APBs will interfere more with secondary tasks of the same modality than tasks of different modalities.

On one level, Hypothesis 3 is well supported by the data. There clearly are interactions effects which change the relationship of APB type and workload, depending on task. However, the expected effect of same-modality interference was not found. We expected the Voice APB to interfere more with the reading and web tasks due to common verbal coding. Instead, depending on the analysis, either the direction of the APB effect is reversed, or the modality effect is not present. This would suggest that the modality effect is weak for these stimuli and tasks. For example, a conversational verbal task instead of a reading task might have produced a more obvious effect of modality interference with the Voice APB.

The objective performance data show a clear learning effect over the course of the experiments. In most cases, the statistical main effects were being driven by the baseline-APB difference, rather than the difference between APB types. Meanwhile a performance increase trend was found over sessions in three of the six performance measures. Despite an attempt to minimize learning effects through the implementation of a break between experimental sessions, participants clearly remembered and improved on the tasks.

The time estimation data show that for the Voice APB condition, participants' estimation of hold time was not significantly different from the actual hold time. Meanwhile, the Electronic APB

condition sees participants significantly underestimating their time in the hold queue by an average of around 20 seconds. As suggested by the time perception literature, people tend to judge more complex stimuli to be shorter in the presence of a secondary task (Hicks et al, 1976). The lack of ISI, novelty factor, and sense of constant progression in the Electronic APB may have increased its complexity enough to produce such an effect.

Implications

While the data show only mixed support for the experimental hypotheses outlined above, the results from this series of studies have practical implications for designers of on-hold telephone stimuli. The data produced from these experiments are significant in a number of ways. The ethnographic data should be of general interest to HCI researchers, practitioners, and communications service providers, as they reveal behavioral patterns in relation to a ubiquitous user interface and situation, albeit with a small slice of the general population. This will benefit future researchers by allowing more realistic settings and tasks to be used, both in the laboratory and in more applied settings.

The workload analysis is also valuable in several ways. As the empirically obtained tasks have proven to be lacking in workload demands in the laboratory, they are evidence that real users are able to choose low-demand secondary tasks in real-life calling situations, such that the performance demand characteristics of APBs is not a critical consideration. Instead, designers may be free to focus on the impact of on-hold stimuli on time estimation, customer satisfaction, and attrition prevention. In high-demand situations where an operator must keep track of multiple tasks, these results could potentially be of use in designing auditory alerts which interfere less with the other tasks. Alerts which are continuous may be more disruptive than those which are intermittent and contains some sort of ISI. The interaction effects are reminders that the relationship between auditory stimuli and multitasking is complex, and designers must carefully model and test their implementations, because unexpected

results can occur.

CONCLUSIONS

This series of experiments was conducted to investigate the effect of two types of auditory progress bar on user workload ratings during multi-tasking with ethnographically derived secondary tasks. As predicted, vocal and electronic APBs affected workload ratings differently depending on the task. However, many predictions based on Multiple Resources Theory were either absent or reversed in the data. Specifically, the electronic APB was rated to have higher mental demand than the voice APB during the reading task, and to have higher overall workload than the voice APB on the math task. Furthermore, the Electronic APB was found to result in caller underestimation of hold time. These results suggest that while the effect of APB type on workload ratings is not in the expected direction, it nevertheless does exist, even if the effect is not strong. In addition, difference between APBs can manifest in any combination of subcomponent scores and overall workload. On the other hand, there was little performance cost associated with either APB. This then implies that for practical APB design, experienced workload may be a secondary factor to caller retention, satisfaction, and task completion.

References

Allport, D.A. (1980). *Attention and performance.* In G.L. Claxton (Ed.) *Cognitive Psychology: New Direction*, 112-153. London: Routledge & Kegan Paul.

Allport, D.A., Antonis, B., Reynolds, P. On the division of attention: A disproof of the single channel hypothesis. *Quarterly Journal of Experimental Psychology, 24*, 255-265.

Anderson, R.J. (1992). Representations and requirements: The value of ethnography in system design. *Human-Computer Interaction, 9, 151-182.*

Angrilli, A., Chrubini, P., Pavese, A., Manfredini, A. (1997). The Influence of affective factors on time perception. *Perception & Psychophysics, 59 (6), 972-982.*

Antonides, G., Verhoef, P. C., van Aalst, M. (2002). Consumer perception and evaluation of waiting time: a field experiment. *Journal of Consumer psychology, 12(3), 193-202.*

Berglund, B., Berglund, U., Ekman, G., Frankenhaeuser, M. (1969). The influence of auditory stimulus intensity on apparent duration. *Scandinavian Journal of Psychology, 10, 21-26.*

Block, R. A., Zakay, A. (1997). Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic Bulletin & Review, 4 (2), 184-197.*

Brigner, W. L. (1988). Perceived duration as a function of pitch. *Perceptual and Motor Skills, 67, 301-302.*

Cameron, M.A., Baker, J., Peterson, M., Braunsberger, K. (2003). The effects of music, wait-length, and mood on a low-cost wait experience. *Journal of Business Research, 56, 421-430.*

Coleman, G.W., Hand, C., Macaulay, C., Newell, A.F. (2005). Approaches to auditory interface design – Lessons from computer games. *Proceedings of the International Conference on Auditory Display, Limerick, Ireland.*

Cooper, G.E., & Harper, R.P. (1969). The use of pilot ratings in the evaluation of aircraft handling qualities (NASA Ames Technical Report NASA TN-D-5153). Moffett Field, CA: NASA Ames Research Center.

Crease, M., Brewster, S. (1998). Making progress with sounds – The design & evaluation of an audio progress bar. *Proceedings of the International Conference on Auditory Display, Glasgow, UK.*

Damos, D.L. (Ed) (1991). *Multiple-task Performance.* Bristol, PA: Taylor & Francis.

Fraisse, P. (1984). Perception and estimation of time. *Annual Review of Psychology, 35, 1-36.*

Gruber, O. (2001). Effects of domain-specific interference on brain activation associated with verbal working memory task performance. *Cerebral Cortex, Nov 2001, 11, 1047-1055.*

Gupta, S., Cummings, L. L. (1986). Perceived speed of time and task affect. *Perceptual and Motor Skills, 63, 971-980.*

Hart, S.G. (2006). Nasa-task load index (NASA-TLX); 20 years later. *Proceedings of the 50th annual meeting of the Human Factors and Ergonomics Society*, 904-908.

Hart, S.G., & Staveland, L.E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*, 239-250. Amsterdam: North Holland Press.

Hazeltine, E., Ruthruff, E., Remington, R.W. (2006). The role of input and output modality pairings in dual-task performance: Evidence for content-dependent central interference. *Cognitive Psychology, 52, 291-345.*

Hicks, R. E., Miller, G. W., Kinsbourne, M. (1976). Prospective and retrospective judgements of time as a function of amount of information processed. *American Journal of Psychology, 89 (8), 719-730.*

Hornik, J. (1984). Subjective vs. objective time measures: a note on the perception of time in consumer behavior. *Journal of Consumer Research, 11, 615-618.*

Hui, M.K., Dube, L., Chebat, J. (1995). The impact of music on consumers' reaction to waiting for services. *Marketing Working Paper Series, MKTG 95.057.*

International Customer Management Institute. (2008). Industry statistics. http://www.icmi.com/statistics/index.aspx?SelectedNode=Statistics. Accessed 10/20/2008.

J.D. Powers & Associates. (2007). Wireless customer care performance study. http://www.wirelessguide.org/comparison/jdpower-customercare06.php. Accessed 12/15/2007.

Jamin, T., Joulia, F., Fontanari, P., Giacomoni, M., Bonnon, M., Vidal, F., & Cremieux, J. (2004). Apnea-Induced Changes in Time Estimation and its Relation to Bradycardia. *Aviation, Space, and Environmental Medicine, Vol. 75 No. 10, 876-880.*

Jex, H.R. (1988). Measuring mental workload: Problems, progress, and promises. In P.A. Hancock & N. Meshkati (Eds.), *Human Mental Workload.* Amsterdam: North Holland Press.

Juslin, P. N., Laukka, P. (2004). Expression, perception, and induction of musical emotions: a review and a questionnaire study of everyday listening. *Journal of New Music Research, 3, 217-238.*

Kellaris, J. J., Cox, A., D. (1989). The effects of background music in advertising: a reassessment. *Journal of consumer research, 16, 113-118.*

Kellaris, J.J., Kent, R.J. (1992). The influence of music on consumers' temporal perceptions: Does time fly when you're having fun? *Journal of Consumer Psychology*, Vol. 1 No.4, pp.365-76.

Klingberg, T. (1998). Concurrent performance of two working memory tasks: Potential mechanisms of

interference. *Cerebral Cortex, Oct/Nov 1998, 8, 593-601.*

Knott, B. A., Kortum, P., Bushey, R. R., & Bias, R. (2004). The effect of music choice and announcement duration on subjective wait time for call center hold queues. *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting, 740 – 744.*

Kortum, P., Ling, A., Su, A., Peres, S.C., & Stallman, K. (2008). Subjective workload assessment for on-hold telephone stimuli. *Human factors and Ergonomics Society Annual Meeting Proceedings, 52*(19), 1517-1521.

Kortum, P., Peres, S.C. (2006). An exploration of the use of complete songs as auditory progress bars. *Proceedings of the 50th Annual Human Factors and Ergonomics Society.* pp 2071-2075, Santa Monica, CA, Human Factors and Ergonomics Society.

Kortum, P., Peres, S.C. (2007). A Survey of Secondary Activities of Telephone Callers Who are Put on Hold. *Proceedings of the Human Factors and Ergonomics Society*, Santa Monica, CA. Human Factors and Ergonomics Society.

Kortum, P., Peres, S. C., Knott, B. & Bushey, B. (2005). The Effects of Auditory Progress Bars on Consumer's Estimation of Telephone Wait Time. *Proceedings of the 49th Annual Human Factors and Ergonomics Society.* pp 628-632, Santa Monica, CA, Human Factors and Ergonomics Society.

Lind, M., & Sundvall, H. (2007). Time estimation as a measure of mental workload. In D. Harris (Ed), *Engineering Psychology and Cognitive Ergonomics,* 369-365. Springer-Verlag Berlin Heidelberg.

Maister, D. H., (2005). The Psychology of Waiting Lines. WWW.davidmaister.com. Retrieved 12/28/2006.

Meyer, J., Shinar, D., Leiser, D. (1990). Time estimation of computer wait message displays. *Proceedings of the Human Factors Society 34th Annual Meeting, 360-364.*

Millen, D.R. (2000). Rapid ethnography: Time deepening strategies for HCI field research. *Proceedings of the 3rd Conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques. 280-286, New York, NY.*

Milliman, Ronald E. (1982). Using background music to affect the behavior of supermarket shoppers. *Journal of mmarketing, 46, 86-91.*

Mishalani, R.G., McCord, M.M. (2006). Passenger wait time perceptions at bus stops: Empirical results and impact on evaluating real-time bus arrival information. *Journal of Public Transportation,* 9 (2), *89-106.*

Munichor, N., Rafaeli, A. (2007). Numbers or apologies? Customer reactions to telephone waiting time fillers. *Journal of Applied Psychology, 92* (2), 511-518.

Nachreiner, F. (1995). Standards for ergonomics principles relating to the design of work systems and to mental workload. *Applied Ergonomics*, *26*(4), 259-263.

Nakajima, Y., Hoopen, G. T., Hilkhuysen, G., Sasaki, T. (1992). Time-shrinking: A discontinuity in the perception of auditory temporal patterns. *Perception & Psychophysics, 51 (5), 504-507.*

Navon, D. (1984). Resources—A theoretical soup stone? *Psychological Review, 91*, 216–234.

Navon, D., & Gopher, D. (1979). On the economy of the human information processing system. *Psychological Review, 86*, 214–255.

Navon, D., & Miller, J. (2002) Queuing or sharing? A critical evaluation of the single-bottleneck notion. *Cognitive Psychology*, *44*, 193-251.

North, A.C., Hargreaves, D.J. (1999). Can music move people? The effects of musical complexity and silence on waiting time. *Environment and behavior, 31 (1), 136-149.*

North, A.C., Hargreaves, D.J., McKendrick, J. (1999). Music and on-hold waiting time. *British Journal of Psychology, 90, 161-164.*

Ozel, S., Larue, J., Dosseville, F. (2004). Effect of arousal on internal clock speed in real action and mental imagery. *Canadian Journal of Experimental Psychology, 58 (3), 196-205.*

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin, 116 (2), 220-244.*

Peres, S.C., Kortum, P. and Stallmann, K. (2007). Auditory Progress Bars: Preference, Performance and Aesthetics. *Proceedings of the International Community for Auditory Display*, ICAD 2007.

Polkosky, M. D., Lewis, J. R., (2002). Effect of ticking rate on user estimation of system response time. *International Journal of Human-Computer Interaction, 14, 423-446.*

Pollack, I. (1953). The information of elementary auditory displays II. *The Journal of the Acoustical Society of America. 25 (4), 765-769.*

Pollack, I. (1953). The information of elementary auditory displays. *The Journal of the Acoustical Society of America. 25 (6), 745-749.*

Pollack, I. (1954). Information of elementary multidimensional auditory displays. *The Journal of the Acoustical Society of America. 26 (2), 154-158.*

Reid, G.B., & Nygren, T.E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In P.A. Hancock & N. Meshkati (Eds.), *Human mental workload*, 185–218. Amsterdam: Elsevier.

Roscoe, A.H. (1987). The practical assessment of pilot workload, AGARD-AG-282. Neuilly Sur Seine, France: Advisory Group for Aerospace Research and Development.

Roscoe, A.H., & Ellis, G.A. (1990). A subjective rating scale assessing pilot workload in flight: A decade of practical use. Royal Aerospace Establishment, Technical Report 90019. Farnborough, UK: Royal Aerospace Establishment.

Roy, M.M., Christenfeld, N.J.S., McKenzie, C.R.M. (2005). Underestimating the duration of future events: Memory incorrectly used or memory bias? *Psychological bulletin, 131 (5), 783-756.*

Rubio, S., Diaz, E., martin, J., & Puente, J.M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA-TLX, and Workload Profile Methods. *Applied psychology: An International Review*, *53*(1), 61-86.

Ruthruff, E., Johnson, J.C., Van Selst, M. (2001). Why practice reduces dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance. 27 (1), 3-21.*

Sawin, D.A., & Scerbo, M.W. (1995). Effects of instruction type and boredom proneness in vigilance: Implications for boredom and workload. *Human Factors*, *37*, 752–765.

Schmutz, P., Heinz, S., Metrailler, Y., & Opwis, K. (2009). Cognitive load in ecommerce applications: measurement and effects on user satisfaction. *Advances in Human-Computer Interaction, 2009.*

Scribner, D.R., Wiley, P.H., Harper, W.H., & Kelley, T.D. (2007). *The effects of workload presented via visual and auditory displays on soldier shooting and secondary task performance.* U.S. Army Research Laboratory, Aberdeen Proving Ground, Human Research and Engineering Directorate, ARL-TR-4224. Accessed 2/3/2010 from http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA471095&Location=U2&doc=GetTRDoc.pdf

Selcon, S.J., Taylor, R.M., Koritsas, E. (1991). Workload or situational awareness?: TLX vs. SART for aerospace systems design evaluation. *Proceedings of the Human Factors Society 35th Annual Meeting*, 62-66. Santa Monica CA: Human Factors Society.

Sivyer, M., Finlay, D. (1982). Perceived duration of auditory sequences. *The journal of General Psychology, 107, 209-217.*

Stallman, K., Peres, S.C., Kortum, P. (2008). Auditory stimulus design: Musically informed. *Proceedings of the 14th International Conference on Auditory Display. Paris, France, June 24-27, 2008.*

Tekman, H. G. (1997). Interactions of perceived intensity, duration, and pitch in pure tone sequences. *Music Perception, Vol 14 No 3, 281-294.*

Tiffin, J., Asher, E.J. (1948). The Purdue pegboard; Norms and studies of reliability and validity. *Journal of Applied Psychology, 32* (3), 234-247.

Thomas, E. A. C., Brown, JR., I. (1974). Time perception and the filled-duration illusion. *Perception &*

Psychophysics, 16 (3), 449-458.

Thomas, E. A. C., Weaver, W. B. (1975). Cognitive processing and time perception. *Perception & psychophysics, 17 (4), 363-367.*

Waard, D.D. (2005). *The measurement of drivers' mental workload.* Ph.D. Dissertation, published by the Traffic Research Center VSC, University of Groningen, the Netherlands. Accessed 1/24/2010 from http://dissertations.ub.rug.nl/FILES/faculties/ppsw/1996/d.de.waard/09_thesis.pdf

Welford, A.T. (1952). The "Psychological Refractory Period" and the timing of high speed performance: A review and a theory. *British Journal of Psychology, 43*, 2-19.

Welford, A. T. (1967). Single-channel operation in the brain. *Acta Psychologica, 27*, 5–22.

Wickens, C.D. (1984a). *Engineering Psychology & Human Performance.* New York: harper & Row.

Wickens, C.D. (1984b). Processing resources in attention. In R. Parasuraman & D. R. Davies (Eds.), *Varieties of attention,* 63–102. New York: Academic Press.

Wickens, C.D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomic Science, 3*(2), *159-177.*

Wickens, C.D., Goh, J., helleberg, J., Horrey, W.J., & Talleur, D.A. (2003). Attentional models of multitask pilot performance using advanced display technology. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 45*(3), 360-380.

Wickens, C.D., Sandry, D., Vidulick, M. (1983) Compatibility and resource competition between modalities of input, output and central processing. *Human Factors, 25*, 227-248.

Young, J.,Reilley, S., Grasha, A.F., Bishop, C., Lis, C., & Roberts, H. (2000). Self-reported stress among individuals in an extended, complex pharmacy verification task. *12th Annual Meeting of APS, Miami, FL.*

Zakay, D., Block, R. A. (2004). Prospective and retrospective duration judgments: an executive-control perspective. *Acta Neurobiologiae Experimentalis, 64, 319-328.*